AL-TR-1991-0118

AD-A248 762

ARMSTRONG LABORATORY

APPLYING NEURAL NETWORKS TO AIR FORCE PERSONNEL ANALYSIS

Vince L. Wiggins

RRC, Incorporated
3833 Texas Avenue, Suite 256
Bryan, TX 77802

Sheree K. Engquist, Captain, USAF
Larry T. Looper

HUMAN RESOURCES DIRECTORATE
MANPOWER AND PERSONNEL RESEARCH DIVISION
Brooks Air Force Base, TX 78235-5000

March 1992

Final Report for Period March 1990 - September 1991

Approved for public release; distribution is unlimited.

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5000

NOTICES

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Office of Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This report has been reviewed and is approved for publication.

SHEREE K. ENGQUIST, Capt, USAF
Project Scientist

WILLIAM E. ALLEY, Technical Director
Manpower and Personnel Research Division

ROGER W. ALFORD, Lt Colonel, USAF
Chief, Manpower and Personnel Research Division

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	March 1992	Final - Mar 90 - Sep 91

4. TITLE AND SUBTITLE

Applying Neural Networks to Air Force Personnel Analysis

5. FUNDING NUMBERS

C - F41689-88-D-0251
PE - 62205F
PR - 7719
TA - 20
WU - 20

6. AUTHOR(S)

Vince L. Wiggins
Sheree K. Engquist
Larry T. Looper

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

RRC, Incorporated
3833 Texas Avenue, Suite 256
Bryan, TX 77802

8. PERFORMING ORGANIZATION REPORT NUMBER

AL-TR-1991-0118

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Armstrong Laboratory
Human Resources Directorate
Manpower and Personnel Research Division
Brooks Air Force Base, TX 78235-5000

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

Armstrong Laboratory Technical Monitor: Capt Sheree K. Engquist, (512) 536-2257.

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The principal objective of this task involved evaluating artificial neural networks for application to personnel modeling by examining areas representative of many personnel models. The four areas chosen were airmen reenlistment, the determinants of reenlistment, and the effects of policy levers; pilot training and more specifically the likelihood of candidates successfully completing Undergraduate Pilot Training; projection of aggregate time series personnel flow rates; and productive capacity of airmen as it relates to aptitude and experience. In addition, the productive capacity analysis was expanded into a working computer prototype allowing the user to examine the effect of changing aptitude/experience mixes on productive capacity. Performance was compared against traditional techniques such as regression analysis.

14. SUBJECT TERMS

Artificial neural networks, Learning vector quantization, Probabilistic networks, Back propagation, Personnel system modeling

15. NUMBER OF PAGES 70

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev 2-89)
Prescribed by ANSI Std 239-18
298-102

CONTENTS

Contents (Continued)

List of Figures

List of Tables

List of Tables (Concluded)

PREFACE

This is the second task in a two-stage effort to assess the potential for applying neural network methodologies to the Air Force personnel field. The research is performed in support of the force management programs of the Human Resources Directorate of the Armstrong Laboratory. Techniques and results developed in this task will serve as analysis and decision tools in the Air Force and OASD force management and policy analysis systems.

Three of the four personnel areas addressed in this research are based on prior modeling and analysis efforts: Stone, Looper, & McGarrity (1990); Stone, Saving, Turner, Looper & Engquist (1991); and Faneuff, Valentine, Stone, Curry, & Hagemann (1990). The cooperation of those researchers in providing background information involving those efforts was essential to the completion of this task. In addition, the authors wish to thank Ms. Kathryn Turner for assistance in preparing and modifying this document, Ms. Phyllis Eddy for proofing, and Mr. Darryl Hand for preparing much of the data.

APPLYING NEURAL NETWORKS TO AIR FORCE PERSONNEL ANALYSIS

SUMMARY

In this task, the performance of neural networks is compared against existing models and traditional estimation techniques in 4 Air Force personnel areas: (1) reenlistment analysis and projection, (2) Undergraduate Pilot Training (UPT) selection, (3) aggregate personnel flow-rate projection, and (4) productive capacity analysis. Some neural network architectures can be viewed as nonlinear estimation techniques which derive the form of the final model directly from the relations found in an estimation or training data set. Several extensions to basic neural network architectures were developed during the task to address the requirements of personnel analysis. Based on out-of-sample projections, the networks were found to perform substantially better than existing models in 2 cases and produced similar results in the other 2 personnel areas.

In projecting individual airmen reenlistment behavior, the network models were superior to probit models across all 5 career fields tested. When projecting over an out-of-sample period, the networks displayed a 35 to 100% improvement in simulation R^2 over the probit models. Similar improvement was found in comparisons on an aggregate model of accession and retention. Neural network models projected a series of "future" flow rates excluded from the estimation sample and the results were much better than ordinary or generalized least squares models (5 to 105% improvement in simulation R^2). In addition, the response surfaces of the neural networks indicated structure in the reenlistment model which is consistent with risk averse behavior but difficult to specify in a standard model. These response surfaces also indicated nonlinear structure which would have a dramatic impact on policy decisions versus those implied by a linear model.

In the areas of UPT selection and productive capacity analysis, the networks performed very similar to the standard regression techniques. In both of these cases, the regression models displayed only moderate statistical significance in-sample and obtained only marginal performance projecting out-of-sample behavior. In these cases, the networks were unable to discover any nonlinear or interacting features which improved significantly upon the regression models. This discovery could be attributable to the weak statistical relation between the independent and dependent variables or the fact that the underlying process being modeled is actually linear over the observed range. Even in these cases, the networks were able to obtain models with performance and response similar to the standard regression models. Overall, the extended network architectures were found to be resistant to marginal data sets and broadly applicable to personnel analysis.

INTRODUCTION

The Human Resources Directorate of the Armstrong Laboratory and others in the personnel management, training, and research areas have applied many modeling and analytic techniques to quantify the decisions, behaviors, and flows observed in personnel systems. In recent years artificial neural network (ANN) techniques have demonstrated some impressive results in modeling other complex systems and in classification tasks (Gorman and Sejnowski, 1988; Shea and Lin, 1989; Surkan and Singleton, 1990; Waibel, 1989; Josin, 1990; and others). A more extensive review of neural network literature on problems similar or related to personnel research is available in Wiggins (1990). The success of ANNs in these areas and their potential for application to personnel modeling lies principally in their ability to automatically detect nonlinear and interacting relations among the inputs and output(s) of a system or observed behavior. Most personnel models require the determination of a relation between a set of inputs (known characteristics or conditions) and a target variable such as a decision, capability, flow, or stock. Traditional analytic techniques require that the form of this relation be specified by an analyst before the empirical estimation of the relationship. Often this form is chosen to be linear by default. ANNs allow more complex relations to be developed directly from observed behaviors of the system or group of individuals under analysis.

The principal objective of this task involves evaluating ANNs for application to personnel modeling by examining 4 areas representative of many personnel models. The first area involves airman reenlistment, the determinants of reenlistment, and the effects of policy levers. The second area involves pilot training and more specifically the likelihood of candidates successfully completing UPT. In the third area, projection of aggregate time series personnel flow rates is examined. The final area addresses the productive capacity of airmen as it relates to aptitude and experience. In addition, the productive capacity analysis has been expanded into a working computer prototype which allows a user to examine the effect on productive capacity of changing aptitude/experience mixes.

To assess the capability of ANNs in each of these areas, the performance of each ANN model was compared against the performance of more traditional techniques such as regression analysis. When possible the techniques were chosen from prior studies in the same area and the same data sets were used. In this manner, the original model can be reconstructed for comparison to the ANN model and both models have access to the same information. In all possible cases the performance of both traditional and ANN models was evaluated both in-sample and out-of-sample (on a set of data or over a time period not covered by the sample on which the model was developed).

While the ANN architectures employed in this research will be introduced, this report will primarily address empirical results and comparisons between ANNs and

other modeling methods. A basic introduction to ANNs with emphasis on personnel problems can be found in Wiggins, Looper, & Engquist (1991a) with other introductory articles also available: Klimasauskas (1988), and Cowan and Sharp (1988). A more detailed presentation and referencing of the ANN methods employed here can be found in Wiggins, Looper, & Engquist (1991b).

AIRMAN REENLISTMENT

The first personnel area examined is the reenlistment decision of first-term airmen. Specifically, given an airman eligible to make a reenlistment decision, the airman's demographic characteristics, Air Force policy, and economic conditions at the time of the decision; what is the likelihood the airman will reenlist. A model capturing this type of decision process serves as the cornerstone of most personnel inventory models (see Carter, Skoller, Perring, and Sakaie, 1988; Michelson and Rydell, 1989; Syllogistics and RRC, 1989; and Stone, Wortman, and Looper, 1989). In addition, this area serves as a very good test bed for the capability of ANNs. As reenlistment has historically been of critical planning importance to the Air Force, it has engendered much research activity: Saving and Stone (1982); Saving, Stone, Looper, and Taylor (1985); Kohler (1988); Carter, Murray, Arguden, Brauner, Abrahamse, Greenberg, and Skoller (1987); and Stone, Looper, and McGarrity (1990). Likewise, many reenlistment efforts have been focused on the other services: Warner and Goldberg (1983); Lakhani, Gilroy, and Capps (1984); Terza and Warren (1986); Lakhani (1987); and Smith, Sylvester, and Villa (1989).

While the reenlistment decision has been heavily researched, virtually all of the models tested have been linear in their input terms. Many researchers have employed logit or probit analysis which imposes a fixed nonlinearity on the output, but still has no inherent flexibility. In a few cases, Stone et, al., (1990) and Carter et. al., (1987), 1 or 2 explicit nonlinear interaction terms were directly introduced into the model to account for unexpected results with strictly linear terms. Still, these terms were minimal changes and the form of the interaction and nonlinearity was pre-specified by the researchers. We hope that the flexible form of the ANN models will capture a more complex mapping from the known characteristics of the airman and the decision environment onto the reenlist/separate decision.

In pattern recognition terminology, analysis of the reenlistment decision is a classification problem. Given observable features (gender, marital status, grade, etc.), which class will an airman fall into (reenlist, separate)? Although not in the personnel decision context, classification problems have been 1 of the most active areas of neural network research. Kohonen (1984), Specht (1988), and Moody and Darken (1988) have developed ANN architectures expressly for the purpose of classification. In addition, the back propagation architecture (Werbos, 1974) has been employed extensively for classification: Odom and Sharda (1990); Kimoto, Asakawa, Yoda, and Takeoka (1990); Atlas, Cole, Conner, El-Sharkawi, Marks, Muthusamy, and Barnard (1990); Leung and Zue (1989); and Denker, Gardner, Graf, Henderson, Howard,

Hubbard, Jackel, Baird, and Guyon (1989). In all of these cases an exemplar is being classified into 2 or more groups based on known information. Viewed from this perspective, the reenlistment decision is fundamentally similar to other classification problems. Based on the success of these researchers, there was reason to believe that ANNs would perform well on the reenlistment decision.

The Reenlistment Model and Data

The reenlistment model chosen to be analyzed is taken from the research of Stone et al. (1990). This model is particularly appropriate for ANN analysis because it retains the inputs as separate components of the pecuniary factors: military compensation, selective reenlistment bonus (SRB), and civilian wages. Many other reenlistment models are based on the Average Cost of Leaving (ACOL) construct which aggregates all pecuniary factors into a single ACOL term (see Warner and Goldberg, 1983). Because the form of this aggregation is fixed, it prevents an ANN from searching for potentially more useful methods of combining the pecuniary factors.

Stone et al. (1990) estimated their model over the January 1975 through March 1982 period and validated the resulting equations over the April 1982 through March 1986 period. Each of the major Air Force Specialties (AFSs) were modeled using a separate probit equation estimated on individual level data for all airmen in an AFS eligible to make a decision during the estimation sample time frame. The resulting probit equations were used to predict the reenlistment decisions of airmen eligible to make decisions over the validation sample time frame. The variables used in their model (and also in the current research) are shown in Table 1. (More detailed explanations of the variables can be found in Saving et al. (1985) with bfor, bpas, atud, and employ2 more fully explained in Stone et al. 1990.) These variables reflect a long-term refinement of the reenlistment model through 2 previous revisions (Saving et al., 1982 and Saving et al., 1985) and the extensive out-of-sample testing performed by Stone et al. These input variables and the functional form reflect a mature model based on many years of research and extensive testing. In this sense, it should provide a stringent benchmark against which ANNs can be compared.

The data used in the current analysis is exactly that used in estimating and testing the Stone et al. model. As described in Saving et al. (1985) and Stone et al. (1990) the primary data consist of records extracted from annual snapshots of the Uniform Airmen Records (UAR), with transition data appended from the Airman Gain Loss (AGL) file. Additional information from Bureau of Labor Statistics and Bureau of the Census tapes was used to derive employment rates and civilian wages.

Stone, Looper, and McGarrity (1990) followed the prior work of Saving et al. and estimated probit equations for each 5-digit AFS ignoring the separation by skill level (effectively a 4-digit AFS). In addition, they estimated models at the more aggregate 2-digit AFS level. For the current exploratory research, the analysis is restricted to three 4-digit career fields and two 2-digit career fields as seen in Table 2.

TABLE 1. INDEPENDENT VARIABLES USED IN THE REENLISTMENT MODELS

Independent Variables	Definitions
Demographic	
dnonwhit	Indicator variable: 1 if non-white airmen, 0 otherwise.
ddep2up	Indicator variable: 1 if 2 or more dependents, 0 otherwise.
dsingle	Indicator variable: 1 if single, 0 otherwise.
dfemale	Indicator variable: 1 if female, 0 otherwise.
Education	
dhsup	Indicator variable: 1 if completed high school or more education, 0 otherwise.
Aptitude	
dafqt12	Indicator variable: 1 if Armed Forces Qualification Test (AFQT) mental category I or II, 0 otherwise.
Pecuniary	
bonus	Sum of SRB payments discounted to the date of the *decision*.
bfor	Bonus forward. Computed by subtracting next month's average SRB from this month's average SRB.
bpas	Bonus past. Computed by subtracting the previous month's average SRB from the current month's.
rmc	Present value of the expected earnings stream from regular military compensation.
cwage	Present value of the expected earnings from an income stream in a civilian job similar to that performed in the AFS being analyzed.
employ	Race and gender specific civilian employment rates.
employ2	The square of employ.
Other	
tafms	Total active federal military service at date of decisions.
atud	Constructed variable to reflect changing attitudes toward the military during and after the Vietnam war. A pure function of time, peaks in 1974 then declines.
dqtr2	Indicator variable: 1 if decision made in the 2nd quarter, 0 otherwise.
dqtr3	Indicator variable: 1 if decision made in the 3rd quarter, 0 otherwise.
dqtr4	Indicator variable: 1 if decision made in the 4th quarter, 0 otherwise.

TABLE 2. AFS CODES EXAMINED IN THE REENLISTMENT ANALYSIS

AFS Code	Description
272X0	Air Traffic Control
316X1	Missile System Maintenance
426X2	Jet Engine Mechanic
30XXX	Communications-Electronics Systems
47XXX	Vehicle Maintenance

NOTE: AFS codes and descriptions are taken from the October 1984 Airman Classification Structure Chart. Any relevant AFS changes over the time frame of the sample have been mapped into or out of these codes.

Modeling Methods

In addition to probit analysis, logit analysis and ordinary least squares were performed during the current work to provide alternate statistical based comparisons. The results from these statistical techniques were compared against 3 neural network architectures: back *propagation*, *probabilistic neural network* (PNN), and learning vector quantization (LVQ). All of the models were trained or estimated on the individual level exemplars or observations from the estimation sample.

While the parametric techniques are better known, the neural network techniques may require a brief introduction. The basic concept of neural networks involves the application of many simple processing elements (neurons) in the solution of a problem or task. While their inspiration and heritage stems from the biological and neurological sciences, the steps to perform an ANN analysis are mathematical. The simple processing elements are deployed into a network architecture which allows communication between the elements. Rules are then used to adapt or train the network to its environment. The rules can implement either self-organization (when the network does not have a specific goal) or supervised training (when the network has a specific goal or set of goals). The organization of the processing elements and the rules which govern them typically define the architecture of an ANN.

Ordinary Least Squares (OLS)

OLS is a frequently used technique in many disciplines and provides a baseline for the other techniques. Despite terminology differences, OLS also provides the same classification results as linear discriminant analysis (Ladd, 1986). When the

dependent variable is binary or dichotomous (as it is in the reenlistment/separation classification problem), the application of OLS is often referred to as the linear probability model. In this context, the output of the linear probability model for a specific airman is interpreted as the probability that the airman will reenlist. This can be seen in Equation 1, which shows the probability of reenlisting for airman *i* as a linear function of fixed coefficients (β_1, β_2, etc.) and the characteristics of airman *i* (*gender, dependents*, etc.). The final component of the equation (e_i) is the difference between the probability predicted by the equation and the actual outcome (1 = reenlist, 0 = separate) for the airman. As is well known, the OLS technique chooses the coefficients such that the sum of squared errors ($\sum e_i^2$) over all candidates in the estimation sample is minimized. Given the covariance matrix of the dependent and independent variables, there is a closed form solution for this minimum sum of squared error coefficients (see Kmenta, 1971 for details). Letting P_i represent the probability candidate *i* will reenlist:

$$P_i = \alpha + \beta_1 gender_i + \beta_2 dependent_i + ... + e_i \qquad (1)$$

or

$$P_i = CX_i + e_i \qquad (2)$$

Where:

C is the vector of coefficients (β_1, ...)

X_i is the vector of inputs for airman *i* ($gender_i$, ...)

The use of OLS on a dichotomous dependent variable (reenlist/separate) poses 2 problems, 1 conceptual and the other technical. The output of the OLS model can vary between negative and positive infinity while the probability it represents is restricted by definition to remain between 0 and 1. The problem, conceptually, is how to interpret a model result (probability) below 0 or above 1. In practice, results below 0 are assigned a probability of 0 and those above 1 are assigned a probability of 1. While this is somewhat troublesome it does not invalidate the use of OLS for dichotomous dependent variables. On a more technical front, OLS can be shown to be inefficient when applied to dichotomous dependent variables (Maddala, 1985). Put simply, the binary nature of the dependent variable violates the OLS efficiency assumption that the regression errors be normally distributed. Again, this does not invalidate the use of OLS in this case; it merely points out that the reported standard errors are larger than the actual standard errors and that all of the information in the sample is not put to best use by the technique.

Logit

Logit analysis addresses both the conceptual and technical problem with the linear probability model (OLS). Logit is based on a maximum likelihood estimation which does not have the same efficiency restrictions as OLS for binary dependent variables. In addition, logit always produces an estimate between 0 and 1 which conforms to the standard conception of probability. The solution of an estimated logit equation has the closed form shown below (using the representations of Equation 2):

$$P_i = \frac{1}{1 + e^{-Cx_i}} \tag{3}$$

The coefficients (C) of the equation are determined by maximizing the likelihood of observing the actual reenlist/separate behaviors of the airman in the estimation sample assuming the cumulative errors follow a logistic distribution (see Maddala, 1985).

Probit

Probit analysis is closely allied with logit; the sole distinction being the assumption of a normal distribution of errors by probit. There is no simple closed form solution for the probability of a probit estimation. The solution requires the integration of the normal probability density function. Saving et al. (1985) and Stone et al. (1990) employed the probit estimator in all of their work. In practice, the 2 techniques produce very similar results and this study will sometimes employ only logit.

Back propagation

Back propagation is the most widely applied neural network architecture developed to date. It is a supervised learning procedure in which the network adapts to the inputs and desired outputs by error correction. While various error measures can be used, the most common (and the 1 used in this study) involves minimizing the sum of squared prediction errors over all of the training exemplars. This is the goal of linear regression. However, in the case of back propagation, several nonlinear processing elements (each having the same form as the logit function shown in Equation 3) are applied to the problem. Use of multiple elements allows the network to "discover" the underlying relationship between the inputs and the outputs. This relationship is not constrained to linearity (as in OLS) and can in fact take on any non-linear form (Hornik, Stinchcomebe, and White, 1989; Funahashi, 1989; Hecht-Nielson, 1987). This freedom to fit the data generally implies that back propagation will require more information (usually more sample observations) than regression techniques to find meaningful relationships. In standard regression analysis, the researcher provides extra information to the model by specifying a fixed underlying functional relationship. Implementation methods and the theoretical development of back propagation within a personnel modeling context are discussed in Wiggins et al.

(1991b). The original development of back propagation can be found in Rumelhart and McClelland (1986) and Werbos (1974).

The freedom of a back propagation model to fit the inputs to the desired output is directly related to the number of processing elements it employs and the number of layers into which they are organized. Typically the complexity of a back propagation solution is constrained by limiting the number of processing elements in the network (Karnin, 1990; Mozer and Smolensky, 1989; Ash, 1989; Sietsma and Dow, 1988). This type of restriction is somewhat related to a specification search using regression techniques. However, instead of imposing a fixed functional form, small numbers of nonlinear processing elements limit the overall flexibility of the trained network. Restrictions of this form are usually designed to enhance the generalization capability (or out-of-sample performance) of a network.

Given the large stochastic component (statistical variation or noise) in most personnel data sets, it is important to limit the complexity of the trained network model. Without some constraint, it is quite likely that a back propagation network will simply "memorize" all of the exemplar results without formulating a model which performs well on individuals or exemplars with new combinations of characteristics. This behavior is similar to the problem of over-fitting a data set using a high degree polynomial and regression analysis.

An alternative to limiting the number of processing elements, is limiting the amount of training time allowed. The back propagation method is adaptive and requires many (often thousands) passes through a data set (epochs) before training is complete. Several researchers (Rumelhart, 1990; and Kimoto et al., 1990) have suggested stopping the training early as a means of improving out-of-sample generalization. Using samples with known properties, Morgan and Bourlard (1990) suggest that both network size and amount of training may be important in determining generalization capability. An example of over-training on actual reenlistment data can be seen in Figure 1. As training proceeds along the epoch axis, both in- and out-of-sample performance improves -- root mean square error (RMSE) declines. However, after a certain point during training, the in-sample performance continues to improve while out-of-sample performance degrades substantially. This portion of the training could be categorized as memorizing the noise in the training sample rather than extracting relevant features from the sample. By watching the network's performance on a hold-out sample on which training is not performed, the training process can be terminated before this memorization process begins.

Stopping training early is the primary method employed in the current research to improve generalization. Improving generalization by choosing the number of processing elements is more of an art than a science and the early stopping methods were found to be much more effective and less ad hoc in personnel analysis. In tests with various network sizes, it was found that relatively small networks were required to capture all of the structure in the personnel models examined in the current research. Networks with 3 to 9 processing elements organized in a network with a single hidden

layer proved sufficient for all analyses. Larger and more complicated networks were unable to perform better than these simple networks.

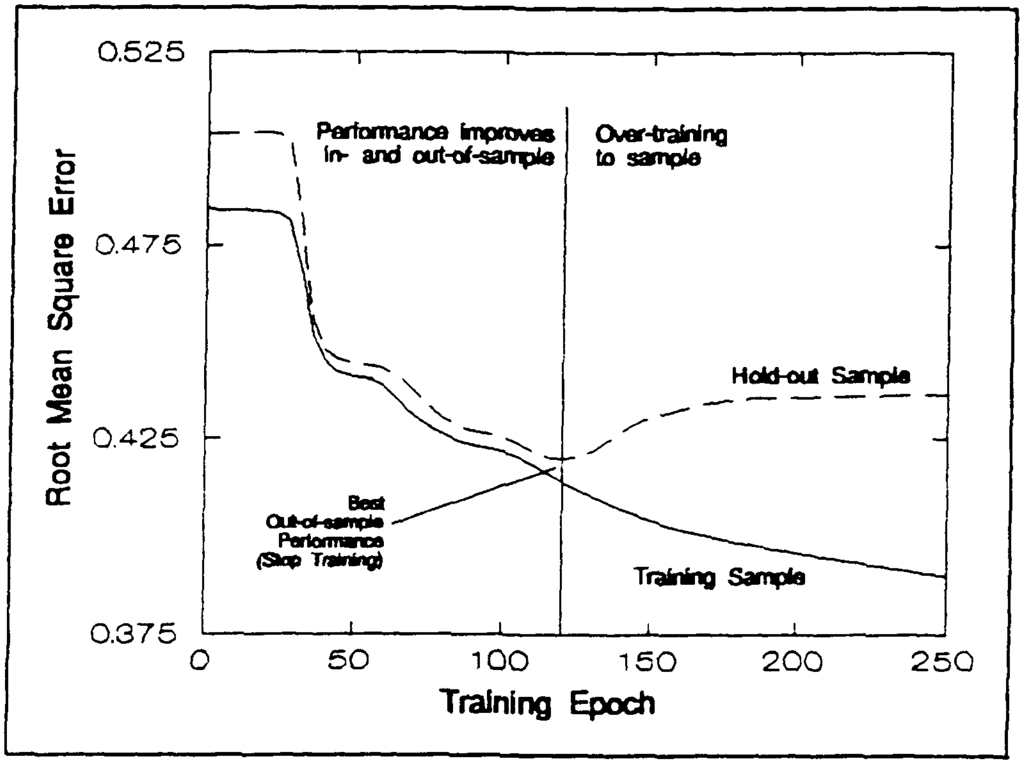

Figure 1. Training path for back propagation. Training sample (solid line) and hold-out sample (dashed line) performance as the number of training passes through the training data set increases.

Probabilistic Neural Network

The PNN is somewhat unusual among neural networks in that it directly implements a version of a more traditional classification technique using neural network concepts. As developed in Specht (1988 and 1990), the PNN forms a separate, nonparametric, probability density function (PDF) for each of the classes or categories to be separated (reenlist or separate for the current problem). Each of the PDF's is multidimensional (as many dimensions as inputs) and by definition the area under the PDF sums to 1. The neural network forms a PDF for a given class, such as reenlisters, in the following manner. The inputs for each candidate in the estimation sample who reenlists are stored in a separate processing element. As shown by Parzen (1962) and Cacoullos (1966), these sample input values can collectively be used to estimate the underlying population PDF for all candidates who will reenlist. The process used in this study forms the PDF from small, multivariate gaussian kernels centered on each airman in the estimation sample. Summing

across these kernels, as shown in Equation 4, provides a point estimate of the probability density for an individual or point in the input space.

$$f_r(X) = \frac{1}{(2\pi)^{p/2}\sigma^p}\,\frac{1}{m}\sum_{e=1}^{N}\exp\left[-\frac{(X-X_{Re})^t(X-X_{Re})}{2\sigma^2}\right] \tag{4}$$

Where:

p is the dimensionality of the input space (i.e., the number of inputs: gender, dependents, etc.).

σ is a "smoothing parameter" which determines the size or extent of the gaussian kernel around each training exemplar.

N Number of training exemplars or observations.

X Vector of inputs at the point for which the density is to be measured (or the vector for a new exemplar to be classified).

X_{Re} Input vector for the reenlister training exemplar e.

t Matrix transpose operator.

While these component distributions are gaussian, the resulting PDF can assume any continuous form. The only adjustable parameter in the PNN is a smoothing factor which determines how smooth the generated PDF will be. As seen in Figure 2, if the smoothing parameter is large, the generated PDF will approach a multivariate normal distribution centered at the input means of all training exemplars in a class (e.g., reenlisters). If it is very small, the PDF will consist of many small, gaussian "bumps" centered at the inputs of each airman in the class. While the smoothing parameter is usually fixed by the researcher, in this study the parameter is allowed to be set based on the amount of noise in the training sample. The parameter is chosen such that the RMSE across all training exemplars is minimized. When the error for each exemplar is computed, it is withheld from the sample so that the estimate of its class membership is based on all other exemplars in the training sample (hold-one-out sampling). In this manner, an optimal (in terms of the RMSE) smoothing parameter is chosen. It is also possible to choose separate input weights using this same methodology such that the effective length of the input space is compressed along some dimensions and accentuated along others. This process can provide for more efficient use of the data if the training sample is small.

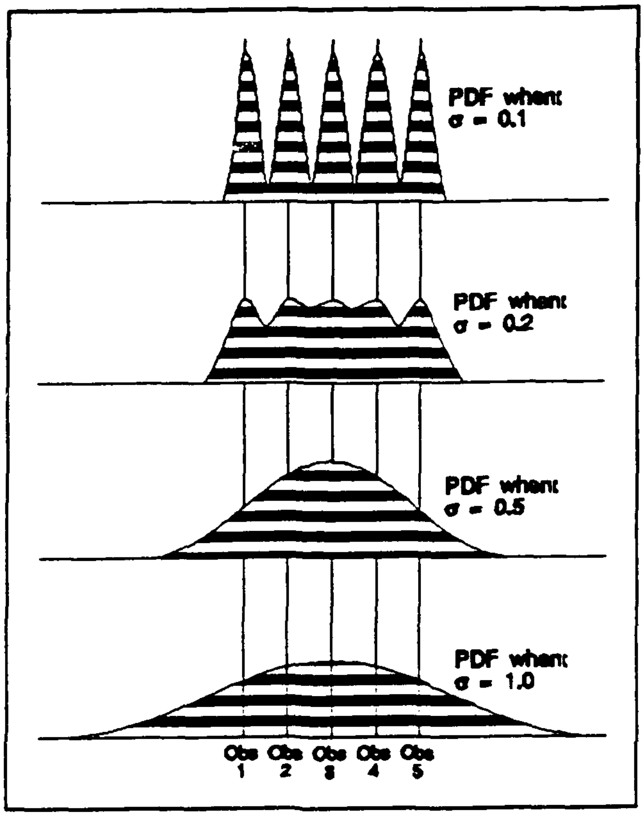

Figure 2. Examples of PNN gaussian kernals. Effect of changing the scaling parameter on the form of an estimated PDF. All 4 PDFs are derived from the same 5 sample observations.

Using the process just outlined, a PDF can be generated for candidates who reenlist and another for those who separate (using the estimation sample to construct the PDFs). Once the PDFs of the 2 classes are known, a simple "Bayes strategy" can be used to determine the most likely class of a new airman. Letting h_r represent the proportion of decision makers reenlisting in the estimation sample, $f_r(X)$ represent the PDF of airmen reenlisting, and $f_s(X)$ represent the PDF of airmen separating (where both PDFs are functions of all airmen's characteristics X); the Bayes rule becomes:

$$\begin{aligned} &\textit{reenlist if:}\ \ h_r f_r(X) > (1-h_r) f_s(X) \\ &\textit{separate if:}\ \ h_r f_r(X) < (1-h_r) f_s(X) \end{aligned} \qquad (5)$$

As stated, this rule assumes the cost of a misclassification is the same whether an airman who actually reenlists is classified as a separator; or, an airman who actually separates is classified as a reenlister. A slight modification to Equation 5

provides for differential costs of misclassifying reenlisters or separators (see Specht, 1990). The probability an airman will reenlist can also be derived from the components of Equation 5. It is merely the ratio of the 2 sides of either inequality.

Learning Vector Quantization

The LVQ technique was developed specifically to solve classification problems (Kohonen, 1984). It operates in a manner similar to a nearest neighbor classifier (Duda and Hart, 1973) in that an unknown candidate is classified according to the behavior of a reference vector. In the simplest nearest neighbor classifiers, an airman from the validation sample is assumed to behave the same as the airman from the estimation sample whose inputs are nearest to his own. This nearness can be measured in many ways, but is usually taken to be the Euclidean distance between the validation airman's and the estimation airman's input vectors.

The LVQ method is somewhat analogous to the PNN except all of the estimation airmen are not retained for comparison with each validation sample member. Instead, a fixed number of reference vectors are allocated and each is assigned to a processing element in the network. Each reference vector is assigned 1 of the 2 (or more) classes (e.g., reenlist/separate). These vectors are then trained to the estimation sample in the following manner. An estimation sample airman is presented to the network and the distance from each of the reference vectors is computed. The nearest reference vector then adapts itself to the candidates inputs. If the vector correctly classifies the airman (it is a "reenlist vector" and the airman reenlists or a "separate vector" and the airman separates), the reference vector moves its weights (reference inputs) toward those of the airman. If the vector incorrectly classifies the airman, the weights are moved away from the airman's input values. After several passes through the data set, a stable set of reference vectors are generated.[1] Airmen from the validation sample are assumed to behave in the same manner as those from the estimation sample who are captured by the same reference vector. Kohonen has shown that this method can arbitrarily approximate complex decision rules by using piecewise linear boundaries (Fig. 3).

Reenlistment Results

Several modeling techniques were tested on the reenlistment data using split sampling methods to validate the models. The modeling techniques included the linear probability model, probit, logit, LVQ, PNN, and several variations of back propagation. Two different sample splits were used to assess the ability of the models to generalize. In the first split, a random sample containing about one quarter of the decision makers was held out during estimation or training. The second split was made according to the period during which an airman was eligible to make a decision.

[1] In work reported here, before this supervised training process begins, the weight vectors are allowed to adapt without comparison to the actual class of the exemplar (reenlist/separate). This provides an initial distribution of reference vectors which mirrors the PDF of the estimation data set (see Kohonen, 1989).

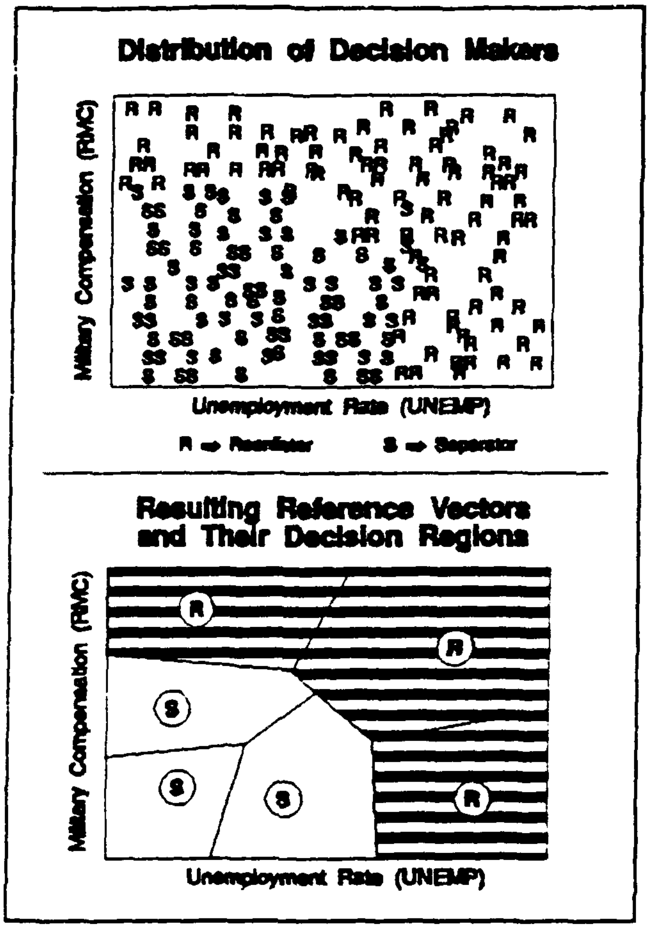

Figure 3. Decision boundaries formed by an LVQ network. A hypothetical distribution of airmen at a reenlistment/separation decision point and the decision regions formed by applying the LVQ architecture to this distribution.

The temporal split used in Stone et al. (1990) was also employed here (January 1975 through March 1982 for the estimation sample, April 1982 through March 1986 for the validation sample). In each case, the models resulting from estimation or training on the estimation sample were used to produce predictions of the decisions of those airman in the hold-out sample.

Performance Measurement

The simulation R^2 was employed to measure the performance of each model's predictions. As seen in Equation 6, the computation of the simulation R^2 directly mirrors the computation of the coefficient of determination (R^2) reported by most regression packages. However, instead of generating the total variation in the validation sample from the validation sample mean reenlistment rate, the mean from the estimation sample is used. This mean is more appropriate for validation samples because one does not know a priori the mean for an unseen sample.

$$\text{Simulation } R^2 = 1 - \left[\frac{\sum_{i=1}^{n} (P_i - A_i)^2}{\sum_{i=1}^{n} (\bar{A}_e - A_i)^2} \right] \qquad (6)$$

Where:

P_i is the predicted reenlistment probability for airman i.

A_i is the actual reenlistment/separation decision for airman i.

$\bar{A}_e$ is the mean reenlistment rate over the estimation sample.

n is the number of observations in the validation sample.

Like the coefficient of determination, the simulation R^2 has an upper bound of 1.0 which is achieved when every decision is perfectly predicted with a probability of 1.0 (e.g., all reenlisters are assigned a predicted reenlistment probability of 1.0). Rarely does the measure approach 1.0 for problems such as reenlistment where the dependent variable is binary. Unlike the coefficient of determination, it is possible for the simulation R^2 to be less than 0. If the modeling method fails to produce a projection which is better than that produced by the in-sample mean reenlistment rate, the simulation R^2 will be negative.

When using neural network techniques, and in particular back propagation, it is important to track out-of-sample performance. Because the back propagation architecture is extremely flexible, it was possible with some network configurations to train a network to have virtually no error on a training sample of decision makers. However, this level of training results in very poor out-of-sample performance.

During testing, the proportion of correct decision predictions were also computed. In this case, the model was forced to produce a definite reenlist/separate decision rather than a probability. Comparisons between the models using this proportion were very similar to comparisons using the simulation R^2.

Variations on Back Propagation

As discussed earlier, it is possible to improve the out-of-sample performance of back propagation networks by stopping training before the network has completely stabilized. The simplest method involves tracking the performance of the network on the actual validation sample during training. The training is stopped when the best performance is achieved on the validation sample (BP Hold in Table 3). While effective, this method utilizes some feedback information from the validation sample

which is unavailable when making an actual projection over some new time horizon or set of airmen. (If the distribution of errors and their correlation to the inputs does not change from the estimation sample to the validation sample, this criticism does not hold. However, in finite samples and particularly when the samples are drawn from different periods, it is very unlikely that these error distributions will meet this criterion.) Because standard regression techniques cannot take advantage of validation sample information, 2 stopping methods which do not employ the validation sample were also employed here. All 3 methods are outlined in Table 3. Note that only the first 2 methods are applicable to non-temporal split samples.

TABLE 3. BACK PROPAGATION TRAINING STOPPING METHODS

Method	Description
BP Hold	Compute the validation sample RMSE after each training pass through the estimation sample. Choose the amount of training which produces the smallest RMSE on the validation sample.
BP Tri-sample	1. Randomly split the original estimation sample into separate pre-estimation and prevalidation samples. (In this case two-thirds of the estimation sample was placed in the preestimation sample and one-third in the prevalidation sample.) 2. Train only on the preestimation sample while tracking the RMSE on the prevalidation and preestimation samples. 3. Save the preestimation RMSE at the training point where the prevalidation RMSE is best. 4. Retrain the network on the original estimation sample (both the preestimation and prevalidation samples). Stop training when the RMSE from the preestimation sample matches the one saved in Step 3.
BP Temporal	1. Split the original estimation sample into separate temporal pre-estimation and prevalidation samples. (In this case the period January 1975 through March 1980 was used in the preestimation sample and April 1980 through March 1982 in the prevalidation sample.) 2. Again, train only on the preestimation sample while tracking the RMSE on the prevalidation and preestimation samples. 3. Save the preestimation RMSE at the training point where the prevalidation RMSE is best. 4. Retrain the network on the original estimation sample (both the preestimation and prevalidation samples). Stop training when the RMSE from the preestimation sample matches the one saved in Step 3.

The BP Tri-sample approach avoids using the actual validation sample by randomly generating its own validation sample from the estimation sample. This pre-validation sample can then be used in a first pass to determine when training should be stopped on the entire estimation sample. This method is applicable to both random and temporal split samples. If a temporal data set contains temporally unstable features which are relevant only during some periods, the BP Tri-sample method may result in over training. Because the prevalidation and preestimation samples span the same period, the network may be allowed to train to features which exist over that period but disappear over the validation period. If there are underlying stable features, the temporal subsampling used in the BP Temporal method may help avoid over training.

Results on Random Samples

The validation performance on randomly selected hold-out samples which span the entire time frame of the data sets are shown in Table 4. This random split-sample measures each model's ability to extract information from an estimation sample that is consistent with the validation sample. In keeping with the work of Saving et al. (1985) and Stone et al. (1990), separate models were developed for each of the 4- and 2-digit AFSs considered.

TABLE 4. VALIDATION SAMPLE RESULTS RANDOMLY SELECTED VALIDATION SAMPLE

	Simulation R^2 by modeling technique				Sample observations	
AFS	Probit	BP Hold	BP Tri-Sample	LVQ	Estimation	Validation
272X0	.158	.311	.322	.272	4,315	1,455
316X1	.041	.120	.068	.046	844	282
426X2	.274	.385	.382	.324	7,170	2,363
30XXX	.153	.311	.306	.233	20,849	6,929
47XXX	.211	.311	.307	.264	3,637	1,151

As can be seen in Table 4, the 3 neural network techniques performed better than probit on all 5 AFSs. The results for OLS and logit were virtually identical to probit and are therefore not reported. While the LVQ method consistently exceeded the out-of-sample performance of probit, the back propagation method (using either sampling approach) provided the best performance in all cases. For the reported LVQ results 1 PE (or reference vector) was allocated for every 30 observations in the AFS's estimation sample. Other numbers of PEs were tested and this rule produced the best results in most cases. Comparing the 2 back propagation approaches on this sample, the extra information gained by tracking the actual validation sample (BP Hold results) did not substantially affect the results. When compared to probit, both back propagation methods projected very well, explaining 40 to 100% more of the variation in an AFS's validation sample.

Results on Temporal Samples

The random split-sample results obscure 2 possible confounding factors when attempting to predict over periods not included in an estimation sample. First, as mentioned earlier, it is possible the decision process contains features which change over time. These features would be important over some periods and irrelevant, less important, or different over other periods. Second, the ranges and variation in the inputs may differ across periods. In this case, a model estimated over 1 period must extrapolate along its response surface when asked to predict results for input ranges outside those in the estimation sample. Both of these factors could substantially affect out-of-sample performance. With flexible model structures, the effect would be similar to over-fitting the sample data. Such a model might consider features no longer present or place too much confidence in the expected range of inputs. To evaluate the impact of these factors, the temporal split-samples employed by Stone et al. (1990) were used and the results are reported in Table 5.

For the temporal split-sample, the LVQ technique was replaced by the PNN. The PNN uses the hold-1-out method discussed earlier on the estimation sample to choose an optimum smoothing parameter. Again, the back propagation methods generally produced the best out-of-sample predictions. As expected, when back propagation was able to track performance on the validation sample (BP Hold), it produced the best projections. However, the temporal subsampling method (BP Temporal) produced comparable results on all AFSs except 316X1. The results of tracking a random estimation subsample (BP) Tri-sample were mixed. For jet engine mechanics (426X2), the performance was actually worse than probit while good projections were obtained for 30XXX. Apparently some of the AFSs have experienced some changes from unmodeled inputs or temporally unstable relations which caused the BP Tri-sample method to over-fit the estimation sample. For the AFSs analyzed, the BP Temporal method appears quite resistant to these problems.

TABLE 5. VALIDATION SAMPLE RESULTS TEMPORAL VALIDATION SAMPLE (APRIL 1982 THROUGH MARCH 1986)

	Simulation R^2 by modeling technique					Sample observations	
AFS	Probit	BP Hold	BP Tri-Sample	BP Temporal	PNN	Estimation	Validation
272X0	.139	.222	.154	.205	.120	3,663	2,107
316X1	-.194	.116	-.173	-.035	-.023	1,010	116
426X2	.269	.368	.141	.365	.173	5,785	3,750
30XXX	.155	.244	.241	.316	.*	18,001	9,777
47XXX	.198	.331	.300	.312	.214	3,144	1,644

*The PNN training could not be completed on AFS 30XXX due to the excessive computations required by the hold-one-out training technique and the large size of the career field.

In the case of 316X1s, the small validation sample enhanced the effectiveness of tracking the sample. Still, the BP Hold result indicated that sufficient information existed in the estimation sample to produce reasonable projections if a proper stopping point was chosen during training. Despite performing worse than the mean estimation reenlistment rate, the BP Temporal method far exceeded the performance of the probit analysis on 316X1.

Overall, the back propagation network performed quite well compared to an established model of Air Force reenlistment. The subsample training stopping heuristics proved critical in improving the performance of back propagation, particularly the BP Temporal method on the temporal split-sample. When large samples are available to serve as training exemplars, back propagation appears to be a viable option for model development.

PILOT TRAINING

In this phase of the research neural network and more standard statistical techniques were applied to the classification of UPT candidates. In particular, UPT candidates were classified on their ability to successfully complete the training program. The principal goal in this phase was to identify successful (and unsuccessful) UPT candidates based on easily obtained information from the Portable

Basic Attributes Test (Porta-BAT) and the Air Force Officer Qualifying Test (AFOQT). As was the case with the reenlistment model, only the binary pass/fail UPT criterion was used to determine candidate success. Grades and other ordinal or continuous measures of success were not explored in this study. Again, the disposition of candidates into pass/fail categories can be viewed as a typical classification problem and all of the techniques employed in the reenlistment problem are applicable.

UPT Data

The primary data for this analysis were based primarily on Porta-BAT results and consists of records for 885 candidates for UPT. The fields in this data set contain the UPT final outcome (pass/fail), the candidate's age, 16 AFOQT subtest scores, and 19 scores and composites from the Porta-BAT. A complete listing of these fields is contained in Table 6 where they are identified by Neuralbat in the source column.

Additional training data (UPT entry date, UPT completion date, courses taken, etc.) was obtained by matching the social security numbers of candidates from the original file against the Flying Training UPT/UNT file in the Air Force Human Resources Laboratory (AFHRL)[2] computer system. All 885 candidates were successfully matched against this file. Several binary variables reflecting the year and quarter the candidate entered UPT were generated from these data elements and were included in some of the analyses. Again, the complete list of fields used is in Table 6.

Finally, each candidate's social security account number was matched against tri-annual snapshots of the Uniform Officer Records (UOR) over the period from the third quarter 1982 through the third quarter 1989 (21 snapshots in all). For each candidate, their first occurring UOR was excerpted and appended to the original data set. Eighty-five of the candidates could not be matched to the UOR, presumably because they "washed out" of UPT between snapshots or returned to reserve units without appearing on the UOR. In fact, 63 of the 85 unmatched candidates were found on Air Force reserve files. Of the 22 remaining unlocated candidates, 18 were UPT failures and the disposition of the remaining 4 could not be determined. UOR data were used to construct binary variables reflecting a candidate's individual and demographic characteristics: gender, number of dependents, education level, etc. The UOR variables used in the analyses are also listed in Table 6.

Several important characteristics of the data set should be noted. First, the Porta-BAT was not given to graduates of the Air Force Academy. Since Porta-BAT was a primary source of data, Academy graduates are excluded from the study. Likewise, many Reserve Officers Training Corps (ROTC) candidates are also excluded because they took the Flight Screening Program (FSP) at their respective colleges. Most of the candidates on the original (Neuralbat) data set were Officer Training School (OTS) graduates with some ROTC graduates from smaller ROTC programs. Table 7 contains a breakdown of the UPT candidates by source of commission.

[2] AFHRL has been redesignated Human Resources Directorate, Armstrong Laboratory.

TABLE 6. DETERMINANTS OF UPT SUCCESS

Variable	Description	Source
AGE	The UPT candidates age	Neuralbat
VA2	AFOQT subtest, verbal analogies	Neuralbat
AR2	AFOQT subtest, arithmetic reasoning	Neuralbat
RC2	AFOQT subtest, reading comprehension	Neuralbat
DI2	AFOQT subtest, data interpretation	Neuralbat
WK2	AFOQT subtest, work knowledge	Neuralbat
MK2	AFOQT subtest, math knowledge	Neuralbat
MC2	AFOQT subtest, mechanical comprehension	Neuralbat
EM2	AFOQT subtest, electrical maze	Neuralbat
SR2	AFOQT subtest, scale reading	Neuralbat
IC2	AFOQT subtest, instrument comprehension	Neuralbat
BC2	AFOQT subtest, block counting	Neuralbat
TR2	AFOQT subtest, table reading	Neuralbat
AI2	AFOQT subtest, aviation information	Neuralbat
RB2	AFOQT subtest, rotated block	Neuralbat
GS2	AFOQT subtest, general science	Neuralbat
HF2	AFOQT subtest, hidden figures	Neuralbat
PS2X1S	Standardized two hand coordination X score	Neuralbat
PS2X2S	Standardized complex coordination X score	Neuralbat
PS2Y2S	Standardized complex coordination Y score	Neuralbat
PS2Z2S	Standardized complex coordination Z score	Neuralbat
ENCRTS	Encoding speed, avg. response time, correct responses	Neuralbat
ENCPERS	Encoding speed, percent correct	Neuralbat
MRTRTS	Mental rotation, avg. response time, correct responses	Neuralbat
MRTPERS	Mental rotation, percent correct	Neuralbat
ITMRTS	Item recognition, avg. response time, correct responses	Neuralbat
ITMPERS	Item recognition, percent correct	Neuralbat
TMSSLPS	Time sharing, slope level of difficulty, min. 3-10, learning rate	Neuralbat
TMSICPS	Time sharing, intercept level of diff., min. 3-10, learning rate	Neuralbat
TMSDIFS	Time sharing, average level of difficulty, min. 11-10	Neuralbat
TMSRTS	Time sharing, average response time, correct responses	Neuralbat
WKARTS	Word knowledge, average response time, dual task	Neuralbat
WKAPERS	Word knowledge, average response time, correct responses	Neuralbat
WKABETS	Word knowledge, percent correct	Neuralbat
AIAHIRS	Activities interest inventory, number of high risk choices	Neuralbat
AIARTS	Activities interest inventory, average response time	Neuralbat
DUPT85	Binary, 1 if candidate entered UPT in 1985	Flytrain

TABLE 6 (CONCLUDED)

Variable	Description	Source
DUPT86	Binary, 1 if candidate entered UPT in 1986	Flytrain
DUPT87	Binary, 1 if candidate entered UPT in 1987	Flytrain
DUPTQTR2	Binary, 1 if candidate entered UPT in the 2nd quarter	Flytrain
DUPTQTR3	Binary, 1 if candidate entered UPT in the 3rd quarter	Flytrain
DUPTQTR4	Binary, 1 if candidate entered UPT in the 4th quarter	Flytrain
DBLACK	Binary, 1 if candidate is black	UOR
DRACEOTH	Binary, 1 if candidate is a non-black minority	UOR
DMARRIED	Binary, 1 if candidate is married	UOR
DDIVORCE	Binary, 1 if candidate is divorced	UOR
DOTSDIST	Binary, 1 if candidate is a distinguished OTS graduate	UOR
DMASTERS	Binary, 1 if candidate has completed a master degree	UOR
DACAD3	Binary, 1 if college major is biological or agricultural science	UOR
DACAD4	Binary, 1 if college major is math	UOR
DACAD6	Binary, 1 if college major is social science	UOR
DACAD9	Binary, 1 if college major is engineering	UOR
DDEP2UP	Binary, 1 if candidate has 2 or more dependents	UOR

Sources: Neuralbat — Original Porta-BAT data set provided by AFHRL/MOEA.
Flytrain — Flying training UPT/UNT file.
UOR — Uniform Officer Records.

TABLE 7. SOURCE OF COMMISSION

Source of Commission	Candidates	Percentage
Officer training school	625	70.6
OTS distinguished graduate	83	9.4
ROTC from UOR field	58	6.6
ROTC from on ROTC files	63	7.1
Unknown from UOR field	34	3.8
Not found on UOR or ROTC	22	2.5
Total	885	100.0

A second restriction in the data involves the year the candidate entered UPT. As seen in Table 8, there is a very uneven distribution of candidates over the sample's 6-year span. As mentioned earlier, these counts are a small proportion of all UPT entrants. A particular anomaly occurred for the 1988 entrants: all 21 failed UPT. In all other years, at least 60% of the entrants passed UPT. Unless there were some mitigating circumstances, the odds of all 21 entrants failing UPT in 1988 is infinitesimally small. For this reason, the 1988 entrants were left out of the analyses.

TABLE 8. UPT ENTRANTS ON THE PORTA-BAT DATA SET BY YEAR

Year	Entrants	Percentage
1982	2	.2
1984	123	13.9
1985	228	25.8
1986	346	39.1
1987	165	18.6
1988	21	2.4

The original Neuralbat data set was divided into 2 nearly equal size samples of 442 and 443 candidates with the intent that the first sample be used to estimate the models and the second sample be held out to validate the models. Each model considered was estimated and validated on those complete samples. When data elements from the UOR were included in an analysis, 102 candidates were dropped from the analysis. Eighty-five candidates were dropped because they could not be found on the UOR and the remaining 1988 entrants (17) were dropped because their observed results were extremely unlikely (as discussed earlier). This decreased the original samples to 396 for the estimation sample and 387 for the validation sample.

All of the continuous variables in the data set (i.e., those whose names do not begin with a d) were standardized to mean 0.0 and standard deviation 1.0 based on the 396 candidates in the estimation sample. This adjustment helped the performance of the LVQ and PNN networks by preventing any variable from dominating the distance computations required by the networks. The adjustment has little or no effect on any of the other techniques tested.

Additional Modeling Method (Stepwise Regression)

Given the large number of potential inputs and the small size of the samples, some method of selecting "important" inputs for a regression would be helpful for the UPT problem. Stepwise regression is a simple, data driven method of performing this

function. While the application of stepwise regression always causes some problems when making inferences from the resulting equation and standard errors (Leamer, 1978), the use of a hold-out or validation sample mitigated these problems in this case.

In a stepwise regression, inputs are either added to or removed from the equation based on some statistical test of their marginal significance. The measure used in this study is the partial-F statistic of the input to be added or dropped from the equation (see Koutsoyiannis, 1977). Several variations exist on the stepwise regression procedure. The method employed here starts with only a constant term and inputs are added 1 at a time if they pass the partial-F test (for implementation details see Computing Resource Center, 1989). In addition, each added input was retested on each pass to insure it still passed the partial-F test. While several significance levels were tested, the result reported here required an input to have a partial-F of 4.0 to enter or remain in the equation. This fairly restrictive level reduced the number of inputs actually used to between 5 and 8.

UPT Empirical Results

Seven modeling techniques were applied to 3 sets of input variables or determinants: (1) all of the variables on the Neuralbat data set; (2) all of the Neuralbat variables, temporal indicators from the Flytrain data set, and candidate characteristic indicators from the UOR (i.e., all variables in Table 6); and (3) a selected set of 8 inputs (listed in the footnote to Table 12). As mentioned in the data section, the second and third sets of variables required reduction of the sample sizes to 396 for the estimation sample and 387 for the validation sample due to candidates who could not be matched to the UORs. The initial goal of this phase was to employ only the Neuralbat data in classifying candidates; it was hoped the addition of temporal and candidate indicators from the Flytrain and UOR files would improve the lackluster performance of the original models. The third set of variables stemmed from the recognition that the limited number of observations would not support the large number of variables in the data set.

The modeling techniques used were basically the same as those applied to the reenlistment problem. Probit was not used and stepwise regression was added for the reasons just discussed. In addition 2 forms of back propagation training were employed: the early stopping method outlined in the reenlistment section and the more traditional training to stability (until the network weights stop changing and the network has stopped adapting). With such a small data set, the stopping criterion used here was always the performance of the actual validation sample.

Porta-BAT and AFOQT Results

Table 9 shows the results of estimating linear probability (OLS) and logit models for all of the AFOQT subtest and Porta-BAT scores (Neuralbat variables) for candidates in the estimation sample. The estimation results from this model support

several early concerns about the application of this small data set to the pass/fail UPT problem. Looking at the t-statistic column of the table for both OLS and logit estimation, only 4 of the 36 (not counting the constant) variables are statistically different from 0.0 at the .05 level of significance. Given the .05 significance level and the 36 variables in the equation, one would expect to find 1.8 (or about 2) significant variables even if each of the variables were generated by uncorrelated random processes. Finding only 4 significant variables is not much better than what would be expected from a data set of purely random noise. There is little reason to place much confidence in this model.

TABLE 9. OLS RESULTS ON NEURALBAT ESTIMATION SAMPLE

	Ordinary Least Squares		Logit	
Variable	Coefficient	t-statistic	Coefficient	t-statistic
AGE	-.001	-0.044	.009	0.068
VA2	.020	0.746	.123	0.870
AR2	-.000	-0.011	-.009	-0.059
RC2	-.023	-0.728	-.135	-0.817
DI2	.001	0.049	.007	0.049
WK2	-.029	-0.841	.150	-0.813
MK2	.038	1.274	.212	1.383
MC2	-.018	-0.642	-.117	-0.811
EM2	.001	0.038	.009	0.071
SR2	.008	0.307	.056	0.416
IC2	.051	2.017*	.277	2.150*
BC2	-.015	-0.591	-.085	-0.632
TR2	.048	1.917	.245	1.935
AI2	.064	2.603*	.343	2.682*
RB2	-.019	-0.767	-.112	-0.856
GS2	-.042	-1.530	-.229	-1.599
HF2	-.034	-1.398	-.186	-1.459
PS2X1S	-.044	-1.618	-.218	-1.612
PS2X2S	-.036	-1.072	-.195	-1.159
PS2Y2S	-.058	-2.119*	-.283	-2.055*
PS2Z2S	.004	0.127	.017	0.106
ENCRTS	.032	1.021	.177	1.084
ENCPERS	.003	0.124	.025	0.187
MRTRTS	.010	0.377	.070	0.499
MRTPERS	.031	1.290	.167	1.366
ITMRTS	-.069	-2.403*	-.368	-2.500*
ITMPERS	.001	0.056	-.000	-0.002
TMSSLPS	-.034	-0.676	-.151	-0.585

TABLE 9 (CONCLUDED)

Variable	Ordinary Least Squares Coefficient	Ordinary Least Squares t-statistic	Logit Coefficient	Logit t-statistic
TMSICPS	-.005	-0.096	.015	0.052
TMSDIFS	.225	0.623	.095	0.506
TMSRTS	-.017	-0.688	-.112	-0.853
WKARTS	.034	1.221	.182	1.261
WKAPERS	.019	0.609	.109	0.671
WKABETS	.004	0.144	.003	0.023
AIAHIRS	-.023	-1.009	-.112	-0.927
AIARTS	-.014	-0.554	-.058	-0.422
CONSTANT	.672	31.429*	.849	7.369*
	Number of obs:	442	Number of obs:	442
	F-test (36, 405):	2.15	Log Likelihood:	241.9
	Prob > F:	0.0002	chi2:	75.54
	R^2:	0.1602	Prob > chi2:	0.0001

*95% probability the coefficient is different from 0 using a 2-tailed Student's t-test.

Several factors contribute to these weak results. First, the dependent variable is a binary (pass/fail) measure of success in UPT. Binary dependent variables do not provide as much information as continuous variables that might measure a level of success, such as UPT grades or performance as a pilot. This finding leads to the second factor -- small sample size. While 442 observations are often more than sufficient with a continuous dependent variable, several thousand observations are often required in the case of binary dependent variables. The third factor is the homogeneous nature of the candidates. All of the candidates on this sample attended UPT and had already successfully completed a Flight Screening Program. They were also required to meet other aptitude profiles. In general, there is very little difference along the input variables between those who pass and those who fail UPT. The first and third factors could be overcome in the absence of the second factor. If sufficient observations about each candidate were available, the binary dependent variable will provide enough feedback for a relationship to be established. The same is true for the homogeneous candidate pool. If there are even tenuous relationships between the input variables and UPT success, they can be found with sufficient examples of the relationship (and a correctly specified model).

Table 10 displays the in- and out-of-sample (estimation sample and validation sample) performance of the parametric and neural network techniques. For these models only the 36 variables from the Porta-BAT and AFOQT listed as Neuralbat in Table 6 were used as inputs. The simulation R^2 measure is again used to compare the performance of the models. As with the reenlistment problem, the validation

sample performance was considered most important. It measures the modeling technique's ability to extract relevant features from the estimation sample and generalize those results to classify candidates with new input vectors. Alternately, the estimation sample R^2 measures the modeling technique's ability to summarize the information in the data provided for estimation or training.

Looking at Table 10, one can see that estimation and validation sample performance did not correlate well for most of the techniques. While the regression techniques and standard back propagation seemed to capture some of the behavior in the estimation sample, this performance did not extend to the validation sample. This poor validation sample performance sustains the earlier conjecture that the regression models might be inadequate given the insignificance of their coefficients. In addition, back propagation (when trained to stability) performed worse than using the mean pass rate from the estimation sample. The modified back propagation (BP Hold, described earlier) performed best out-of-sample with LVQ a somewhat distant second and stepwise regression showing at least some ability out-of-sample. In general, the in-sample performance of the 3 network models was less misleading than the regression based models (with the exception of stepwise regression). Still, none of the methods used performed well on the validation sample.

TABLE 10. ESTIMATION AND VALIDATION SAMPLE PERFORMANCE ON 36 PORTA-BAT AND AFOQT VARIABLES

	R^2	
Modeling Technique	Estimation Sample	Validation Sample
Linear Probability Model (OLS)	.163	.008
Logit	.167	.004
Stepwise regressions[1,2]	.104	.017
Back propagation, trained to stability	.436	-.253
Back propagation, BP Hold	.116	.054
Learning vector quantization (LVQ)	.063	.021
Probabilistic neural network (PNN)	.059	.000

[1] Uses the forward stepwise model and requires partial F-value > 4 for a variable to remain in the model.

[2] Final stepwise variables: TR2, AI2, PS2X1S, PS2Y2S, ITMRTS.

The worst validation performance was obtained by the model which best fit the estimation sample data -- back propagation trained to stability. Given the flexibility of the back propagation method, this result is not surprising. Even with the simple network architecture employed (only 4 processing elements), the network was still able to generate a model which captured much of the information in the estimation

sample (.436 R^2). Tests using slightly more complicated architectures (12 to 21 processing elements) showed that back propagation could obtain an estimation sample R^2 of .98 to .99. However, in the UPT case, the validation performance decreased in direct proportion to the estimation sample performance. As discussed earlier, this result stems from the ability of a highly flexible architecture to "memorize" the noise or stochastic components in the estimation sample to the detriment of its ability to generalize. This behavior can also be seen in the OLS regression model. With 36 inputs, the model is able to obtain a .163 R^2. However, the model is virtually useless outside the estimation sample (.001 R^2). While the linear model cannot change the form of the relationships, it can misidentify the linear impact of inputs based on the stochastic components of the estimation sample. Fortunately, with regression techniques, the standard errors of the coefficients give a good indication of the ability to generalize. However, the overall equation F-test is a weak test of significance (it merely requires that any coefficient in the model to statistically different from 0). In many cases (including the UPT problem) the F-test is not a good indicator of out-of-sample performance.

Stepwise regression was introduced as a simple selection technique to combat the multitude of input variables and tendency to over-fit with such a small data set. While stepwise performed better out-of-sample than the other regression techniques, the improvement over the linear probability model was minimal. In keeping with the earlier discussion, the estimation sample performance declined as inputs were removed from the model. The stepwise results are based on a 4.0 partial-F criterion which is very stringent and excludes most of the variables. Less restrictive partial-F tests were employed, but decreased the ability of the stepwise model to perform on the validation sample.

Early stopping of back propagation training proved the most effective technique for out-of-sample prediction. While the validation performance was still relatively poor, it was substantially better than any of the other models tested. As seen in the reenlistment results, stopping the back propagation training early helped the network to capture only those features from the data which were useful for generalization. As expected, estimation sample performance was much lower than unconstrained back propagation training but was much more indicative of validation performance.

Porta-BAT, AFOQT, UOR, and Training Results

The poor performance of all the models on the original set of variables possibly reflected the absence of some important determinants of UPT success. By matching the candidates to the UOR, several demographic and educational characteristics were determined. In addition, a match to the flying training data sets allowed generation of annual and quarterly indicator variables to account (in a simple manner) for institutional changes over the time period. Collectively, and with the original Neuralbat data set, this produced 53 variables for each candidate (all listed in Table 6). Table 11 shows the results of applying the 7 modeling techniques to these variables. Almost every technique performed better on this data set, both in-

(estimation) and out-of-sample (validation). Aside from the first back propagation and the PNN each method displayed increased estimation and validation R^2. Still, with the exception of back propagation with early stopping and stepwise regression, estimation sample performance is not indicative of predictive power.

TABLE 11. ESTIMATION AND VALIDATION SAMPLE PERFORMANCE ON ALL 53 VARIABLES

	R^2	
Modeling Technique	Estimation Sample	Validation Sample
Linear Probability Model (OLS)	.242	.025
Logit	.253	.017
Stepwise regressions[1,2]	.149	.050
Back propagation, trained to stability	.686	-.352
Back propagation, BP Hold	.165	.071
Learning vector quantization (LVQ)	.080	.013
Probabilistic neural network (PNN)	.047	.013

[1] Uses the forward stepwise model and requires partial F-value > 4 for a variable to remain in the model.

[2] Final stepwise variables: WK2, IC2, PS2Y2S, DUPT86, DBLACK, DMARRIED, DOTSDIST, DACAD4.

The modified back propagation network continued to perform best out-of-sample and stepwise regression improved most on the validation sample. As can be seen in Table 11, stepwise chose 5 of its 8 variables from the UOR matched variables which were not available on the initial data set. Despite the continuing mediocre performance of even the best models, all of the empirical results indicated that some of the additional variables were important determinants of UPT success. Apparently, the modified back propagation method is best at extracting relevant information from the numerous inputs.

Results with Selected Variables

The large number of inputs was a concern for all of the models discussed so far. The somewhat superior performance of stepwise regression and modified back propagation indicated that a reduction in the number of variables might provide models that perform better out-of-sample. Eight variables were selected from the 53 available based on their consistent performance in several logit models. These 8 variables were then used to form models using all of the techniques except stepwise regression (with a reduced model already selected, stepwise was unnecessary). As seen in Table 12, all of the techniques performed better on the validation sample than they had with the 2 larger sets of inputs. The linear probability model, logit, and modified back propagation had identical validation performance. Even so, it was difficult to demonstrate that any of the models performed better than the mean pass

rate from the estimation sample. A test of the validation RMSE between the logit model and the estimation sample mean showed no difference at the .10 significance level (Steel & Torrie, 1960).

TABLE 12. ESTIMATION AND VALIDATION SAMPLE PERFORMANCE ON 8 SELECTED VARIABLES*

Modeling Technique	R^2 Estimation Sample	R^2 Validation Sample
Linear Probability Model (OLS)	.133	.079
Logit	.141	.079
Back propagation, trained to stability	.253	.021
Back propagation, BP Hold	.161	.079
Learning vector quantization (LVQ)	.080	.004
Probabilistic neural network (PNN)	.059	.067

*Variables used: WK2, IC[illegible] [illegible]X2S, ITMRTS, DUPT86, DMARRIED, DOTSDIST, DACAD4.

Several other models were estimated (or trained) and validated. In some of these models, groups of the AFOQT and Porta-BAT test scores were aggregated. As few as 2 aggregate variables were tried as input in some models. In addition, the entire data set was resampled to produce an estimation sample of 632 candidates and a validation sample of 151 candidates. Many of the techniques were attempted on these samples. The results of these variables and sample choices produced models whose performance was similar to those already reported. None of these models were superior to those estimated on the hand picked variables used in the last set of models.

Performance of Back Propagation

Despite the somewhat weak results of all the models, a comparison between the best models from Table 12 and the modified back propagation model in Table 11 demonstrates an interesting result. By stopping the back propagation training early when using all 53 variables, this method was able to nearly equal the validation performance of the other methods on the best "hand-picked" set of variables. This set of variables is a potentially useful facility when approaching a problem where the relations between the determinants and output variable(s) are difficult to establish.

Given that back propagation networks were initialized to a random starting point before training and the dynamics of back propagation training are not at all well understood, it is difficult to say if this performance is repeatable. While the question remains open, theoretically, an empirical examination of many networks using the

UPT data indicates some interesting results. Table 13 shows the results of applying various back propagation architectures to the 8 and 53 input UPT data sets. As seen in the hidden processing elements (PEs) column, the number of hidden neurons ranged from 0 to 18. A hidden PE arrangement of 9,9 indicates a network with 2 layers of hidden elements each containing 9 PEs. Each PE in the first hidden layer is connected by an individual weight to each input. Each PE in the second layer is connected by a separate weight to each PE in the first hidden layer. The 9 PEs in the second layer are connected to a single output PE which produces the probability of UPT success. Likewise a 6,6,6 arrangement utilizes 3 hidden layers with 6 PEs in each layer.

In addition to changing the number and arrangement of PEs in each network, 2 different training rates were used. As seen in the last column, all of the networks produced virtually indistinguishable results on the data sets with 8 selected inputs. More importantly, all of the networks produced very similar results using all 53 input variables (except the network with no hidden units which essentially implements an adaptive version of logit analysis). In addition, the validation performance was very similar between the networks using 8 inputs and those using 53 inputs. Despite the limitations of the UPT data set, the modified back propagation method was able to seek a model which performs as well as any of the models from the hand-selected data.

TABLE 13. STABILITY OF BACK PROPAGATION PERFORMANCE USING VALIDATION SAMPLE RMSE AS A TRAINING STOPPING CRITERION

Number of Inputs	Hidden Processing Elements	Learning Rate	Training Epochs	Validation Sample Simulation R^2
8 (selected)	0	.01	77	.075
8 (selected)	3	.10	164	.079
8 (selected)	3	.01	640	.079
8 (selected)	18	.10	99	.079
8 (selected)	18	.01	855	.079
8 (selected)	9,9	.01	2,351	.079
8 (selected)	6,6,6	.10	589	.075
53 (all)	0	.01	12	.058
53 (all)	3	.10	11	.071
53 (all)	3	.01	114	.075
53 (all)	18	.10	8	.068
53 (all)	18	.01	83	.071
53 (all)	9,9	.01	560	.075
53 (all)	6,6,6	.10	305	.079
Logit on 8 selected inputs				.079

UPT Summary

The pass/fail classification of UPT candidates posed particularly difficult problems. The candidates had very similar characteristics and those with nearly identical characteristics often had different outcomes. Very fine distinctions were required among essentially similar candidates to determine a single output (pass/fail). While some of the possible determinants displayed the ability to form some distinction among the candidates, the separations were tenuous. This sample was particularly problematic for most of the neural network techniques which tried to discover and establish nonlinear relationships between the input and outputs. There appears to be too little information to conclusively establish even linear relationships. The current sample appears insufficiently large to establish distinctions among similar candidates. Some objective or subjective measure of eventual pilot performance, grades during UPT, or even the binary outcomes from further training would provide more information when trying to distinguish the best UPT candidates. This additional criterion information would assist in forming relationships even with the limited number of observations available from the Porta-BAT.

One encouraging aspect of this phase was the performance of the modified back propagation network when training was stopped early. This method was able to develop good models from an extensive list of variables which, based on the other results, appears to include superfluous and relatively unimportant factors. The modified back propagation method was able to produce models which performed as well out-of-sample as the best hand selected models using all of the 53 available input variables.

AGGREGATE ACCESSION AND RETENTION

A third area addressed in this task involves the estimation and projection of aggregate time-series personnel flow rates of the enlisted corps. As mentioned in the second section, these rates often serve as components of inventory flow models. On an aggregate level, the Air Force personnel system has 3 major flows: non-prior service (NPS) accessions, prior service (PS) accessions, and separations. Separations can be further broken down into voluntary separations by term of enlistment, involuntary separations, and retirements.

Of the aggregate flow rates, NPS accession has received the most attention from researchers (e.g., Ash, Udis, and McNown, 1983; DeVany, Saving, & Shugart, 1978; DeVany and Saving, 1982). However; Stone, Saving, Turner, Looper & Engquist (1991) considered a more complete aggregate model. As with prior research on individual reenlistment, all of these aggregate models employed regression techniques and structural relations which could be made linear in the regression inputs.

Aggregate Time-series Model and Data

The Stone et al. (1991) model included more aggregate flows as dependent variables and served as the basis for developing neural network models. In addition, this model was extensively tested over out-of-sample periods and proved far superior to the rather poor accession results obtained by Ash et al. As described in Table 14, the model includes 4 dependent or output variables: NPS accession rate, PS accession rate, first-term reenlistment rate, and second-term reenlistment rate. (The breakdown of first- and second-term reenlistment rates comprises a necessary disaggregation of the model because these rates reflect fundamentally different underlying decisions.) The model is structural in the sense that each dependent variable has an equation with a specified form and set of independent variables. Theoretical background on the selection of the dependent variables and form of the structural equations is provided in Stone et al. (1991).

TABLE 14. AGGREGATE ACCESSION RETENTION MODEL DEPENDENT VARIABLES

Variable	Definition
NPSRT	Non-prior service (NPS) accession rate (with respect to the 16- to 19-year-old population.
PSRT	Prior service (PS) accession rate (with respect to total population of eligible separators).
RELRT1	First-term reenlistment rate (with respect to eligible-to-reenlist first-term airmen).
RELRT2	Second-term reenlistment rate (with respect to eligible-to-reenlist-second-term airmen).

The model lacks 2 rates required to make it internally complete for aggregate inventory simulation. It does not address term extension rates and reenlistment eligibility rates required by the model itself to develop some of its inputs. Much of the burden for maintaining the eligibility rates would normally fall to the dynamic simulation portion of an inventory model and not the estimation portion. In addition, the researchers were evaluating system estimators, not developing an inventory model. Career-term reenlistment rates and retirement rates are also not considered. Still, the model considers 4 of the principle flow rates and provides far more ground for comparison than the aggregate accession models.

The specific form of the 4 equations estimated by Stone et al. is shown in Table 15. The researchers employed 2 regression based techniques to estimate the structural form of the model, ordinary least squares (OLS) and generalized least

squares (GLS). The OLS estimator was applied to each equation separately while the specific GLS estimator employed allowed for cross correlation among the errors from all 4 equations.

Stone et al. estimated the equations over 1 period and validated their performance over 2 periods -- the period directly preceding the estimation period and the period directly after the estimation period. All estimations and validations were performed on a series of monthly data developed from the Historical Airman Data (HAD) base, military enlistment processing station (MEPS) files, and Bureau of Labor (BLS) sources. The models were estimated on the monthly data spanning the October 1979 through September 1987 period (96 observations). One validation sample consisted of the 9 monthly observations from January 1979 through October 1979. The second validation sample ran on the 12 observations from October 1987 through September 1988 - fiscal year (FY) 1988. For the purposes of this study, the OLS results were reproduced to verify the data set and the GLS results are taken directly from Stone et al. (1991).

Neural Network Approach

The back propagation architecture described earlier was applied to the monthly data just described. The principal method involved creating a separate network for each of the 4 aggregate flow rates considered by Stone et al. Given the minimal differences between the out-of-sample capabilities of GLS and OLS found by the prior researchers separate networks seemed appropriate. Each network employed the inputs from the appropriate equation. For example, first-term reenlistment used RLEMP1, RLWR1, RECR, PSGOAL, QTR1, QTR3, and QTR4 as inputs to the network. In addition, extensive testing was performed with joint networks using all 4 flow rates as outputs and all independent variables as inputs. However, due to the differing training requirements (length of training) of the 4 outputs, these networks did not produce stable results for all of the outputs.

Following the work of the previous researchers, the networks were trained over the October 1979 through September 1987 period (FY 79 - 87). Out-of-sample projections were then made over the 2 validation periods from the previous work. In all cases, the out-of-sample performance of the methods was used to compare results and the simulation R^2 described earlier served as the primary metric for measuring performance.

As discussed earlier, back propagation is capable of over-training networks to the extent that their out-of-sample performance deteriorates That discussion extends to the current model. With sufficient training, back propagation networks with only a few neurons were capable of reproducing the estimation sample (FY 79 - FY 87) with almost no error. However, the out-of-sample performance of these networks was very poor. (Comparison of in-sample performance between the highly flexible networks and regression techniques would be unfair and fruitless.) As with the individual reenlistment problem, heuristics were employed with the time series data during

TABLE 15. AGGREGATE ACCESSION AND RETENTION MODEL EQUATION SPECIFICATION AND INDEPENDENT VARIABLES

Variable	Definition
	Equation 1: NPS Accession Rate (NPSRT)
QUAL	Ratio of AFQT category 1-2 accessions to category 3-8 accessions.
WAIT	Average time spent in the Delayed Enlistment Program (DEP).
EMP	Age specific civilian non-institutional employment rate.
WR	Relative military wage to age specific civilian wage.
RECR	Number of Air Force production recruiters.
FLGOAL	Ratio of current month's force level to Fiscal Year force level goal.
NPSGOAL	Ratio of monthly accession rate to the rate required to meet NPS accession goal.
	Equation 2: PS Accession Rate (PSRT)
PSEMP	Age specific civilian non-institutional employment rate.
RLWR1	Relative military wage to age specific civilian wage.
RECR	Number of Air Force production recruiters.
PSGOAL	Ratio of monthly prior accession rate to the rate required to meet PS accession goal.
	Equation 3: First-term Reenlistment Rate (RELRT1)
RLEMP1	Age specific civilian non-institutional employment rate.
RLWR1	Relative military wage to age specific civilian wage.
DECM1	Ratio of eligible to ineligible first-term airmen.
EOUTS1	Number of first-term early outs.
	Equation 4: Second-term reenlistment rate (RELRT2)
RLEMP1	Age specific civilian non-institutional employment rate.
RLWR2	Relative military wage to age specific civilian wage.
DECM2	Ratio of eligible to ineligible second-term airmen.
EOUTS2	Number of second-term early outs.
	Independent variables in all equations
QTR1	Indicator: 1 in 1st FY quarter, 0 otherwise
QTR3	Indicator: 1 in 3rd FY quarter, 0 otherwise
QTR4	Indicator: 1 in 4th FY quarter, 0 otherwise

training to stop the process before excessive over-fitting could occur. Table 16 outlines the 3 methods used for stopping the back propagation training.

TABLE 16. TRAINING STOPPING METHODS FOR TIME SERIES DATA

Method	Description
BP (79 hold-out)	Choose the amount of training that produces the best out-of-sample performance on the January 1979 through September 1979 (most of FY 1979) sample.
BP (88 hold-out)	Choose the amount of training that produces the best out-of-sample performance on the October 1987 through September 1988 (FY 1988) sample.
BP (inflections)	Stop training at the second negative to positive inflection in the RMSE of the in-sample training path. No information outside of the training sample used.

As can be seen in Table 16, 2 of the methods rely on additional information from outside the estimation sample to determine when training has concluded. The BP (79 hold-out) and BP (88 hold-out) methods monitor the performance over 1 of the 2 validation samples, and select the amount of training which optimizes performance over the monitored sample. This sample is the only information gained from the selected validation sample and no training is performed on the observations from either validation sample. When out-of-sample validation metrics are being computed on the same sample as the monitoring process (e.g., both monitoring and validating the FY 88 sample), this is directly analogous with the BP Hold process employed in the individual reenlistment problem. This is a best case scenario; it is the best that the network being trained can perform on the validation sample given the data in the estimations sample. No point in the training path can perform better out-of-sample. When the opposite validation sample is monitored while computing metrics on 1 validation sample (e.g., monitor FY 88 while validating January 1979 through September 1979), the method is closer to the BP Tri-sample method without the additional training. In this case, no information is obtained over the validation sample being used to validate the out-of-sample performance.

The third training heuristic, BP (inflection), utilizes no information from outside the training sample. For this problem, it was felt the training sample was too small to support a further split during any phase of training. The BP (inflection) method does not split the estimation sample as required by the BP Temporal methods used earlier. Rather it makes use of an empirical observation about the training path of back propagation made by Rumelhart (1990). Specifically, the best out-of-sample performance typically appears near an inflection point in the training path. When the second derivative of the in-sample RMSE with respect to the training epoch switches

from negative to positive an inflection has occurred. While the dynamics of back propagation training are not well understood, this co-occurrence of inflection with good generalization was common enough to warrant examination in this context. The specific inflection point used in these analyses is the second negative to positive occurrence. This inflection point is the 1 most commonly aligned with best out-of-sample performance. Examination of many networks has indicated that the first inflection usually occurs at the point where linear relationships have been established and very often the network mirrors OLS results when examined at this point. The second negative to positive inflection is usually associated with the "discovery" of nonlinear features in the sample.

As a further note on the back propagation architecture used in this analysis, a different transfer function was used by the processing elements in the network. Instead of the sigmoid function used earlier, a hyperbolic tangent function was used in its place (Fahlman, 1988). The hyperbolic tangent is just a symmetric version of the sigmoid ranging from -1 to +1. Work on the time series data and productivity data (discussed later) showed that networks with hyperbolic tangents could be more consistently trained to obtain similar results with similar training epochs. The hyperbolic transfer function required scaling of the output variables between -1 and 1. This linear transformation has no effect on the reported simulation R^2. In addition, all of the inputs to the neural networks were scaled to lie between -1 and 1 using the same transformations applied to the output variables.

Empirical Results on Aggregate Time Series

A comparison of the out-of-sample performance of the 2 regression techniques and 3 variations on back propagation are presented in Table 17. In general, all of the models performed very well on the 1979 validation sample. NPS accessions proved to be the most difficult rate to project, but every model was able to explain more than 50% of the variation in the NPS accession rate. Despite the ability to explain the overall level of all rates, predicting changes in the rates was more elusive. Neither of the regression based projections could be shown to be correlated with actual NPS accessions or PS accessions at the .05 significance level (however all projections were correlated at the .10 level). The BP (79 hold-out) and BP (88 hold-out) PS accession rate projections were correlated with the actual rates at the .05 level. All reenlistment projections were correlated with the actual rates at the .05 level or better. As reported in Stone et al. (1991), little difference could be found between the 2 regression based techniques across any of the rate projections.

Using this validation sample, the neural networks were clearly superior in projecting only 2 of the 4 rates -- PS accessions and first-term reenlistment. In the case of NPS accessions, back propagation could perform better than the 2 regression techniques, but only by "peeking" at its performance on the validation sample -- technique BP (79 hold-out). In all cases, the networks which monitored performance of the validation sample performed best. While this monitoring cannot be performed in

practice when the validation sample is truly unknown, it provides an upper bound on the performance of back propagation on the problem.

TABLE 17. VALIDATION SAMPLE PERFORMANCE (JANUARY 1979 THROUGH SEPTEMBER 1979)

	Simulation R^2			
Modeling Technique	NPS Accession Rate	PS Accession Rate	First-term Reenlistment Rate	Second-term Reenlistment Rate
Ordinary Least Squares	.522	.828	.848	.988
Generalized Least Squares	.540	.797	.853	.988
BP (79 hold-out)	.552	.926	.966	.982
BP (88 hold-out)	.512	.905	.923	.982
BP (inflection)	.506	.831	.912	.950

Based on validation sample projections, the BP *(inflection) method had a* tendency to stop training too early. In particular for PS accession rates and second-term reenlistment rates, the other 2 stopping methods trained over 100 times longer than the BP (inflection) method. Overall, the BP (inflection) method displayed the worst performance among the neural network techniques.

The actual projections of the OLS equation and a back propagation method (BP inflection) are shown in Figure 4. While the BP (inflection) results are the worst on reenlistment of the 3 networks, it provides a model which can be applied to both validation samples without having capitalized on any information for the validation sample. The OLS projection captures the major turning points for the period better than the back propagation projection; however, the OLS projection is biased downward by about 10%.

The 1979 validation sample of 9 observations is rather small, and 1 is not often asked to project the past. The comparison of the methods was extended to the FY 88 validation sample. In this case, the same networks and regression models used to produce the projections for Table 17 were utilized for the FY 88 period to produce the results shown in Table 18.

For this latter period, the improvement of the neural network techniques over the regression methods was quite striking. With the exception of NPS accessions, the BP (79 hold-out) and BP (88 hold-out) models explained more than twice the out-of-sample variations as either OLS or GLS. Two of the 3 BP methods also performed slightly better on the NPS accession rate. Although not typically as strong as the other

2 BP training methods, BP (inflection) outperformed the regression techniques in all cases except OLS on second-term reenlistment.

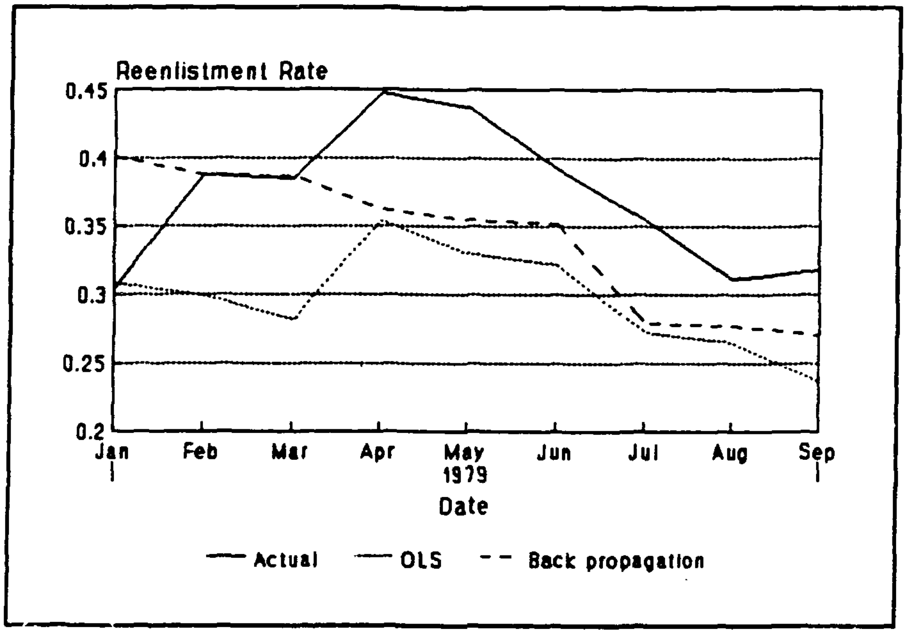

Figure 4. Actual and out-of-sample projections of first-term reenlistment rates for January 1979 through September 1979, ordinary least squares and BP (inflection) models.

Mirroring the 1979 validation sample results, neither regression technique produced accession rate projections (NPS or PS) which were correlated with the actuals at the .05 level of significance (although the NPS projections were correlated at the .10 level). While no network projections were correlated with NPS accessions beyond the .05 level, all network projections of PS accessions, first-term reenlistment, and second-term reenlistment were highly (well beyond .05) correlated with their appropriate actual rates.

Figure 5 displays the FY 88 out-of-sample projections of OLS and BP (inflection). While both project well, the OLS projection misses the upswing in reenlistment by a month, the downturn by 2 months, and projects rates in excess of 100% for 2 months. The back propagation projection captures both the onset and downturn in the reenlistment rate quite accurately.

Projections over the entire estimation and validation sample frames are presented in Figure 6 for OLS and Figure 7 for BP (inflection). As seen in the figures, the BP model has smaller bias over most periods and more accurately reflects the turning points in the reenlistment rate. In particular, the network is better at projecting the rapid swings in the rate.

TABLE 18. VALIDATION SAMPLE PERFORMANCE (OCTOBER 1987 THROUGH SEPTEMBER 1988)

Modeling Technique	NPS Accession Rate	PS Accession Rate	First-term Reenlistment Rate	Second-term Reenlistment Rate
	Simulation R^2			
Ordinary Least Squares	.618	.378	.288	.569
Generalized Least Squares	.606	.317	.237	.323
BP (79 hold-out)	.487	.633	.683	.736
BP (88 hold-out)	.647	.633	.774	.736
BP (inflection)	.644	.550	.772	.436

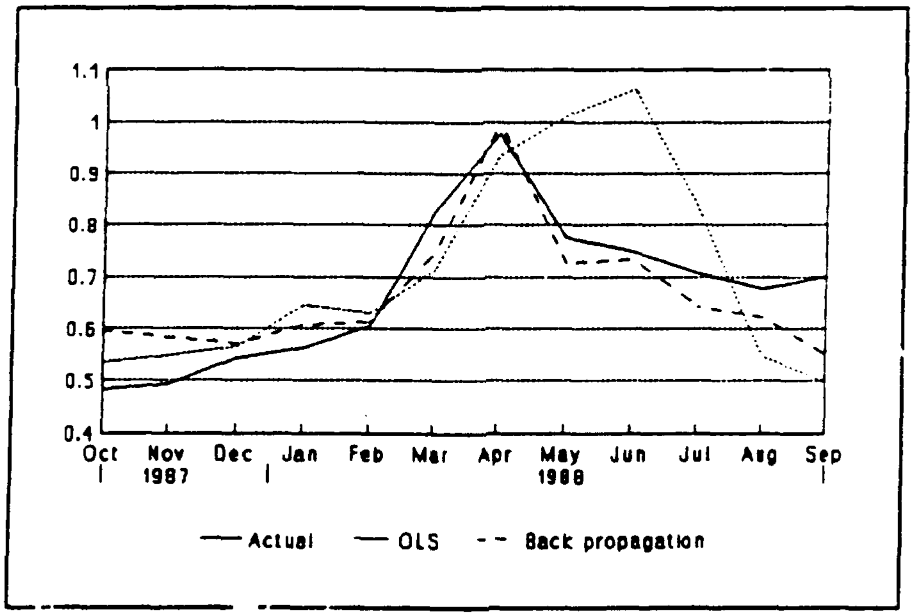

Figure 5. Actual and out-of-sample projections of first-term reenlistment rates for October 1987 through September 1988, OLS and BP (inflection) models.

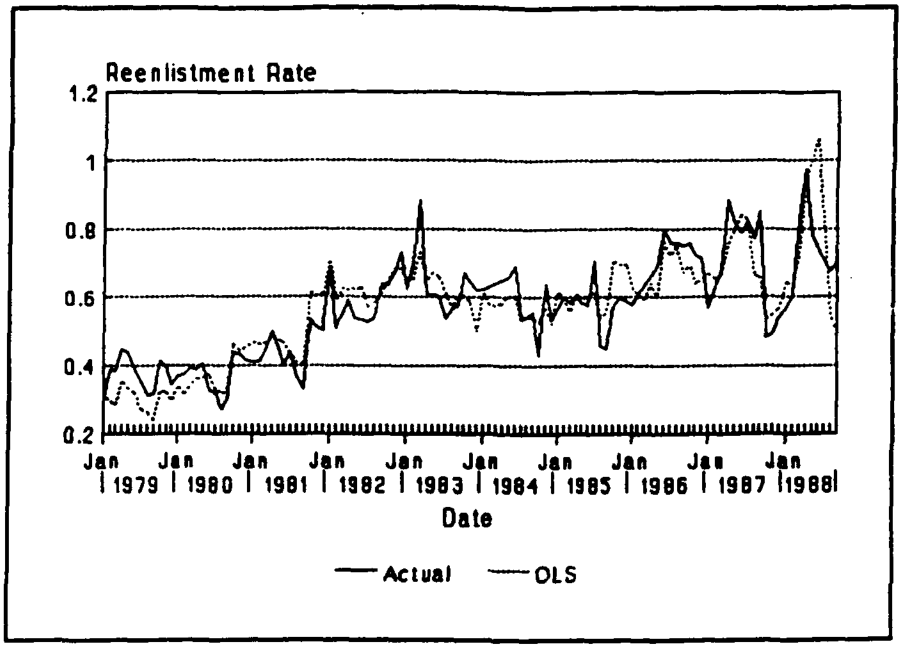

Figure 6. In- and out-of sample simulation of the first-term reenlistment rate using the OLS model.

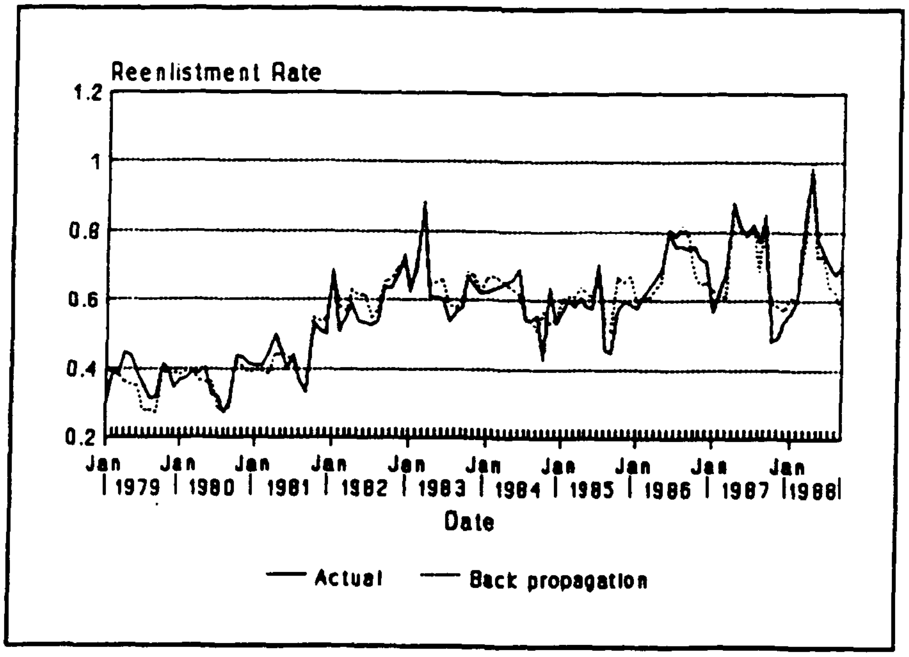

Figure 7. In- and out-of-sample simulation of the first-term reenlistment rate using the neural network model BP (inflection).

Neural Network Reenlistment Response Surfaces

Given the ability demonstrated by back propagation networks in out-of-sample projections, it is interesting to analyze the factors which set the networks apart from the regression techniques. In particular, the networks must be capable of capturing relationships between the independent variables and aggregate rates not specified in the regression models. Two of the principal inputs in each rate equation are a measure of the civilian employment level and relative military to civilian wage. In fact, other than the number of recruiters, most of the other independent variables primarily capture temporal fluctuations in the system which affect distance from goals on the accession side and the content of the decision making pool on the reenlistment side.

The impacts of employment and relative wages on each of the aggregate rates, as modeled by neural networks, are presented in Figures 8 through 11. As outlined in Table 15, each of the relative wages and employment rates was specific to the age of the relevant group for each for the aggregate rate (e.g., 18- to 23-year olds for accessions). An exception to this is the employment rate which uses the sample employment measure for first- and second-term airmen. In addition, the employment rate was converted to an unemployment rate to make the relaticns easier to visualize.

The impact of unemployment levels and relative wages on first-term reenlistment is displayed in Figure 8. To allow this impact to be primarily decision-maker driven, the other 2 independent variables were set to levels which would allow the modeled pool to retain most of the eligible decision makers. Specifically, the ratio of eligible to ineligible first-term airmen (DECM1) was set to its highest value obtained over the sample time frame. This number put the largest proportion of first-term airmen in the eligible decision maker pool. Conversely, the number of first-term early outs was set to 0. Early outs reflect negative decision makers who are no longer in the pool, i.e. their decision is not included in the denominator of the reenlistment rate. The values of the 3 quarterly indicators were set to their mean values over the entire sample.

The figure displays 2 nonlinear but essentially noninteracting impacts. Looking strictly along the unemployment axis, there are 2 relatively flat surfaces where changes in unemployment have little effect on the reenlistment rate. These surfaces occur below 6% unemployment and above 8.5% unemployment. Increases in unemployment above 8.5% do not substantially affect reenlistment; likewise, decreases below 6% have almost no impact. As modeled by the network, the greatest impact of unemployment on first-term reenlistment is between the 6 and 8.5% levels of unemployment.

The relation between relative wages and first-term reenlistment is also nonlinear but of a different form. When military compensation exceeds the civilian wage by less than 10%, changes which keep the relative wage below that level have virtually no effect. As relative wages move from 1.1 to 1.3 level, the effect of a given change in relative wage produces steadily larger changes in the reenlistment rate. Beyond the 1.3 level, a given change in relative wage has a high but constant impact on first-term reenlistment.

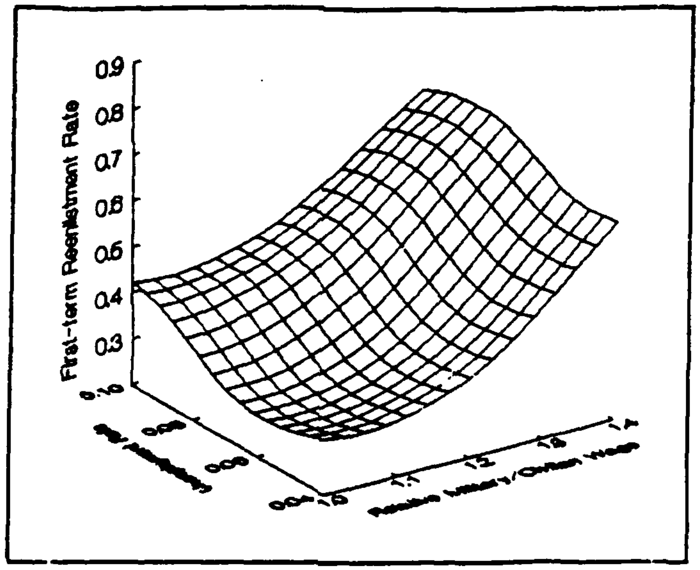

Figure 8. Response of first-term reenlistment rate to unemployment levels and relative military to civilian wage, estimated by the BP (inflect) neural network model.

In addition, it can be seen that the impacts of the 2 factors do not interact. The relation between relative wages and reenlistment is unchanged by shifts in the unemployment rate. While higher unemployment shifts the relation between relative wages and reenlistment up, it does not affect the form. All of the civilian wage impact lines are basically parallel.

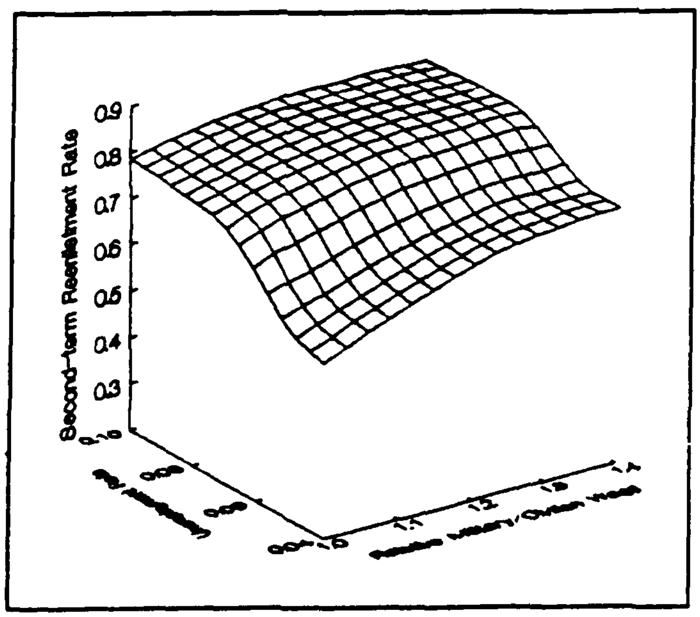

Figure 9. Response of second-term reenlistment rate to unemployment levels and relative military to civilian wage, estimated by the BP (hold-out 88) neural network.

Figure 9 presents the network modeled response of second-term reenlistment to unemployment and relative wage (Fig. 9 is kept on the same scale as Fig. 8 to facilitate comparison). As with first-term reenlistment, the eligible ratio and number of second term early outs are set to the maximum and 0 respectively. A soft threshold phenomenon can again be seen relating reenlistment and unemployment. Below 5% and especially above 7.5% unemployment, changes in the unemployment rate have minimal effect on second-term reenlistment. Again the greatest impact of the civilian unemployment rate is expressed over a 2.5% range in the unemployment level. For second-term reenlistment, the range has shifted down 1% from the transition range observed for first-term reenlistment. This shift reflects an increased risk-aversion exhibited by the older group. As expected, and supported by other research (Saving et al., 1982), the reenlistment rate for second-term decision makers is consistently high and relatively unaffected by changes in military compensation.

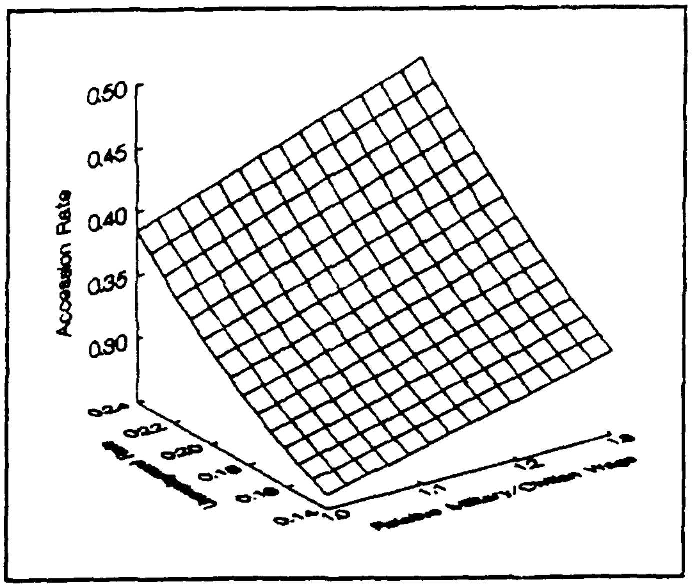

Figure 10. Response of the NPS accession rate to unemployment levels and relative military to civilian wage, estimated by the BP (inflect) neural network. (large values on the unemployment scale reflect the high unemployment rate for the youth population).

Figure 10 displays the impact of unemployment and relative military to civilian wages on NPS accessions. For the purposes of this graph, the other 5 dependent variables and the 3 quarterly indicators were set to their mean values over the entire sample. The graph displays 2 linear, noninteracting but important impacts from the 2 variables. This result is to be expected given the relative performance of the neural network and regressions models. Of the 4 modeled rates, the out-of-sample results were most similar for NPS accessions. Essentially, the neural network has reinforced the original modeler's implicit assumption that no nonlinear features were present in the NPS accessions model.

The modeled response of PS accessions to the levels of the same 2 independent variables is considered in Figure 11. As with NPS accessions, the values of the other variables were fixed at their sample means. Unlike the prior figures, this figure displays considerable interaction between unemployment rate and relative wage in determining PS accession rates. The unemployment level has a dramatic impact on how potential PS accessions respond to changes in relative military to civilian wages. As can be seen in the figure, when unemployment is very low, changes in military compensation have little effect until the military wage exceeds its civilian counterpart by over 20%. However, with high unemployment, the impact of military compensation begins before the relative difference is 10%. In addition, the impact of changing military compensation is much larger and increases faster at low relative wages and high unemployment rates. This is precisely the type of behavior one would expect from a labor group already entrenched in the work-force. High relative wages and changes in those relative wages have much less effect on those who already hold jobs.

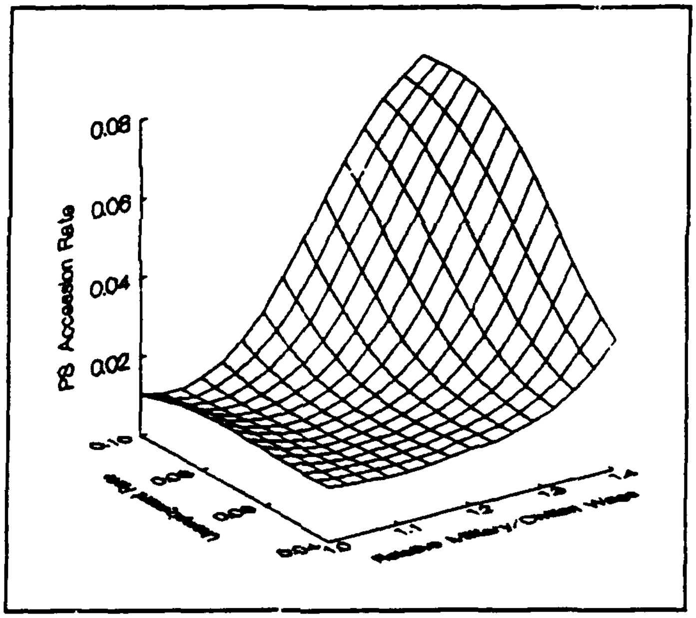

Figure 11. Response of the PS accession rate to unemployment levels and relative military to civilian wage, estimated by the BP (inflect) neural network.

Aggregate Time-series Summary

The out-of-sample performance of the neural network models when projecting aggregate personnel flow rates was quite impressive. In particular, the networks performed much better on the 12 months in the FY 1988 validation sample. A method of stopping the back propagation training was essential to this performance.

Preliminary networks developed without these methods invariably performed very well on the training sample and very poorly on the validation samples.

An examination of some of the response surfaces generated by the network model indicates where improvements over a linear specification were "discovered" by the network architecture. Some of these nonlinear and interacting features were in stark contrast to the linear assumptions made in regression analysis. Most of these features were poorly approximated by the constant effects constraint of linear models or the constant elasticity of log-log models. Although the network was relatively unconstrained in its ability to fit the training data, the features developed were well behaved and extrapolate smoothly. This network model was also in contrast to the nonlinearities generated when high-degree polynomial estimates are used to fit nonlinear surfaces. In most applications, polynomials consistently exhibited strong and unpredictable swings outside the boundaries of the estimation sample. In each case, the nonlinear and interacting features "postulated" by the network model were extremely plausible and often more intuitively appealing than constant or constant elasticity effects over the entire range of an input variable.

A common complaint among researchers modeling time series data involves changes in model structure. When an equation is estimated over one period, its coefficients may substantially differ from those obtained over a different period. A "change in structure" is usually blamed for these differences; however, a glance at Figure 8 will show that a linear model estimated over a period of high unemployment would produce a substantially different result than one estimated over a period of moderate unemployment. A model estimated over both periods would produce a linear average between the two. While this is typically considered a change in structure over time and is the bane of effective projection, the neural network model suggests an alternate interpretation. The model structure has remained constant; it merely contains a richer, more nonlinear structure, than the original estimator was capable of capturing. When networks can capture some of this richer structure, they can be expected to perform significantly better than regression techniques.

PRODUCTIVE CAPACITY

The final area examined in this task was the productive capacity of airmen in the enlisted force. More specifically relations were sought between Air Force experience, aptitude, and productive capacity. This relation serves as a major component in several recently developed models for allocation of personnel (Faneuff, Valentine, Stone, Curry, and Hageman (1990); Stone, Turner, Fast, Curry, Looper and Engquist, 1991). While these researchers focused on the aggregation of productive capacity over time and its allocation effects, the emphasis in this study was determination of productive capacity at any point during active duty service. Any model which produces this result can serve as input to the Faneuff et al. and Stone et al. aggregation and allocation models.

Productive Capacity Model and Data

The specific model of productive capacity examined was taken primarily from Fauneff et al. (1990) and based on the prior work of Carpenter, Monaco, O'Mara, and Teachout (1989). Experience was measured by months of total active federal military service (TAFMS). Aptitude was measured using the subtest scores from the Armed Services Vocational Aptitude Battery (ASVAB) (see Table 19 for a listing of scores). The raw ASVAB subtest scores were rebased to norms from the 1980 Youth Population and standardized to a mean of 50 and a standard deviation of 10.

TABLE 19. ASVAB SUBTESTS

Subtest Mnemonic	Subtest Name
GS	General Science
AR	Arithmetic Reasoning
WK	Word Knowledge
PC	Paragraph Comprehension
NO	Numerical Operations
CS	Coding Speed
AS	Auto Shop Information
MK	Mathematics Knowledge
MC	Mechanical Comprehension
EI	Electronics Information

The Air Force normally employs 4 composites of these 10 subtests when evaluating recruits: Mechanical (M), Administrative (A), General (G), and Electronic (E). These composites (see Table 20) are collectively referred to as the MAGE scores. Admission to each career field is currently based on performance on one or two of these MAGE composites and an overall composite designated the AFQT.

TABLE 20. AIR FORCE ASVAB COMPOSITES

Mnemonic	Composite Name	Composite Computation
M	Mechanical	MC + GS + 2AS
A	Administrative	NO + CS + WK + PC
G	General	WK + PC + AR
E	Electronic	AR + MK + EI + GS
AFQT	Armed Forces Qualification Test	2(WK + PC) + AR + MK

Prior research has focused on these composites as measures of aptitude. The ability to evaluate the effect of changes in the construction of the composite scores would allow much broader latitude for policy analysis. If more effective composite scores or more selective criterion could be developed it would have considerable implications for personnel allocation. To evaluate this problem, models must be developed which relate individual subtest scores to performance or productive capacity.

Productive capacity was measured using objective Walk-through Performance Test (WTPT) measures (Hedge and Teachout, 1986). The WTPT includes hands-on measures involving the observation of airmen actually performing tasks and interview measures which evaluate task knowledge. The measure used in Stone et al. (1991) and employed in the current research is a composite of hands-on and interview test scores -- total WTPT score (TWTPT). The separate hands-on and interview scores were also analyzed and found to behave similar to the TWTPT score. In addition, various supervisor ratings were evaluated, but also found to produce little difference in the results. WTPT data has been gathered on 8 Air Force career fields, 6 of which had been completed at the time of this research. As seen in Table 21, the 6 career fields span all 4 MAGE composites.

TABLE 21. CAREER FIELDS WITH WALK-THROUGH PERFORMANCE TEST DATA

AFS Code	Career Field Title	Composition for Admission	Useful Observations
122X0	Aircrew Life Support	G	176
272X0	Air Traffic Control	G	174
328X0	Avionic Communications	E	68
423X5	Aerospace Ground Equipment	M,E	235
426X2	Jet Engine Mechanic	M	201
492X1	Information Systems Operator	A	201

Following the work of Faneuff et al., the TWTPT score was normalized in each career field to a base considered to be a fully productive airman. For this research, a fully productive person was defined as the median TWTPT score for airmen from an AFS with between 37 and 48 months of service. This median score served as a basis for computing the productive capacity of all other airman in an AFS as shown in Equation 7.

$$P_i = \frac{T_i}{\overline{T}} \tag{7}$$

Where:

P_i is productive capacity for airman i

T_i is the TWTPT score for airman i

$\overline{T}$ is the median TWTPT score for airmen in an AFS with 37 to 48 months of service

Both prior research efforts in this area used a single MAGE score and TAFMS as the independent variables of OLS regressions with productive capacity as the dependent variable[3]. Various functional forms have been employed to estimate the productive capacity function. Carpenter et al., used a logistic function while Faneuff et al., found a linear form with a log TAFMS term to best fit the productive capacity data. Linear, logistic, and log-linear forms were employed in the current analysis as a basis for comparison to network results.

Productive Capacity Results

Four different regression models were estimated for each of the career fields considered: OLS with linear input terms, OLS with log input terms, and two logistic regressions. The logistic regression suggested by Carpenter et al. requires a nonlinear transformation of the dependent variable to obtain an S-shaped relation between the independent variables and the output variable. This functional form is defined only over the region between 0 and 1 for the dependent variable and is not invariant under linear transformations of the variable. In one of the logistic regressions, productive capacity was rescaled to lie between .02 and .98 before applying the logistic transformation. In the other regression, the productive capacity was simply divided by a constant such that the maximum value obtained before the transformation was .95, with the lower bound allowed to fall in proportion to the constant. Two variations on the back propagation architecture were employed; the BP Hold and BP Inflection methods discussed earlier. Due to the relatively small sample size available in each AFS, the more complicated split-sampling approaches were inappropriate.

Two different sets of inputs (or independent variables) were tested for each model on each AFS. To replicate the Carpenter et al. and Stone et al. (1991) models, only the relevant MAGE score for accession selection and job placement was used (in

[3]Sometimes skill level was also included in the regressions.

conjunction with TAFMS). In a second series of models, all 10 ASVAB subtests were entered as inputs to the model with TAFMS.

Each of the AFS samples was randomly divided into an estimation sample containing two-thirds of the observations and a validation sample consisting of the remaining one-third. The models were estimated or trained on the estimation sample with out-of-sample performance based on the validation sample. The simulation R^2 for the performance of each model on the validation sample is presented in Table 22.

TABLE 22. OUT-OF-SAMPLE SIMULATION R^2 FOR PRODUCTIVE CAPACITY MODELS

Modeling Technique	Air Force Specialty Code (AFSC)					
	122X0	272X0	328X0	423X5	426X2	492X1
Model with only the admissions MAGE composite and TAFMS as inputs*						
OLS, all linear terms	.114	.057	.235	.139	.136	.154
OLS, log input terms	.138	.075	.281	.143	.122	.218
Logistic (.02 to .98)	.111	.055	.148	.092	.122	.167
Logistic (X to .95)	.101	.049	.241	.150	.099	.216
BP Hold	.076	.064	.299	.164	.125	.176
BP Inflection	.073	.053	.259	.158	.125	.176
Models using all ASVAB subtests and TAFMS as inputs						
OLS, all linear terms	.039	.077	.465	.127	.086	.110
OLS, log input terms	.064	.125	.457	.127	.090	.194
Logistic (.02 to .95)	-.054	.038	.393	.092	.015	.026
Logistic (X to .95)	.000	.078	.430	.131	.054	.132
BP Hold	.085	.105	.487	.176	.128	.155
BP Inflection	.052	.078	.477	.150	.084	.058

*For AFS 423X5, only the Mechanical (M) composite is used in the first set of models.

No clear pattern emerges from these results which would indicate a superior method of modeling the productive capacity function. It is unclear from these results whether the addition of all subtest scores significantly improved a model's predictive performance. Only the 328X0 results using all subtests was significantly different from the much simpler models using a single MAGE score to represent aptitude. The only model to consistently perform well on most of the AFSs was the BP Hold neural network which was best or second best in all cases except 122X0. (In this case it was still the best of the models developed using the ten ASVAB subtests.) As was

demonstrated in the UPT analysis, the neural networks appear to be able to extract relevant information from small samples with large numbers of input variables. The mediocre performance of the BP Inflection method indicates that finding appropriate stopping points for back propagation poses particular problems with small samples.

The regression models were inconsistent when comparing results between the MAGE and subtest inputs, yet the BP Hold network consistently performed better when given more aptitude information (with the exception of AFS 492X1). Within this context, the BP Hold performance suggests that additional structure is present if appropriate training stopping points can be determined. The subtest models estimated on the WTPT data lack the strength to be applicable in their current form to provide a basis for evaluating new composite scores. However, the consistent BP Hold results indicate that additional data combined with nonlinear analysis might provide a more detailed understanding of the interplay between aptitude, experience, and productive capacity.

CONCLUSIONS

During the course of this task neural networks were compared with traditional estimation techniques and existing models in 4 areas of the Air Force personnel system. In all cases, comparisons among the models were made on the basis of performance over periods or of individuals which were excluded from the samples used to develop the models. This stringent criterion accounts for the inherent ability of neural networks to perform well in-sample.

In 2 of the areas analyzed, the reenlistment of airmen and the projection of aggregate personnel flow rates, the neural network techniques displayed distinct and substantial improvements over existing models. Using the simulation R^2 as a criterion, the improvement was sometimes two- or three-fold over the existing model on the groups tested. These 2 areas offered very different problem domains: a time-series analysis of continuous rates with relatively small samples; and, a dichotomous decision problem with extensive data available. In both cases, the ability of neural networks to derive nonlinear features from the set of training observations proved crucial in the network's superior performance. All of the techniques used in the UPT pass rate and productive capacity research were hindered by the limited and homogeneous nature of the data samples. In both cases, the samples available were small and the individuals in the samples had been previously screened by existing selection criteria. These 2 samples offered much more tenuous examples relating the input factors to the modeled behavior. With these 2 problems, the networks performed as well as the other methods tested and were able to perform better when provided no guidance from the researcher (in the form of selecting specific variables for inclusion or deletion from the analysis). In this sense, the networks performed very well as "model seekers" when confronted with less than ideal data.

Overall, neural networks have demonstrated the ability to significantly improve on the performance of some existing models. This ability is directly related to the amount of nonlinear or complex structure in the system being estimated. A critical concern to anyone conducting research on personnel or other highly stochastic systems is to prevent over-fitting of data. The heuristics employed in this research proved very successful at stopping training before the network was able to generalize outside the estimation. Prevention of over-fitting is an area which has received limited attention in the literature and many refinements are possible. In spite of the extremely successful results obtained in some areas of this study, care must be taken to avoid over-training the networks.

The results on individual reenlistment indicate that any future work in that area should consider the use of back propagation or one of its variants as a modeling technique. The reenlistment problem has shown itself to contain significant structure which is not captured by the current regression based techniques, but is amenable to being modeled with neural networks. Substantial benefits in the ability to evaluate the impact of changes in policy or economic conditions would result from the more detailed relations captured by the networks. Likewise, the results on aggregate rate estimation were extremely encouraging. The model developed by Stone et al. (1991) had already exhibited very good out-of-sample performance. The additional structure realized in the network models proved important for both the projection and analysis of the underlying impact of the factors contributing to the rates. Because of the richer modeling environment offered by neural networks they should be considered for many problems where sufficient data exists to extract relations between known factors and observed behaviors.

Most of the work performed during this research centered on testing the validity of neural networks for personnel data analysis. This work primarily involved testing the performance of trained networks under new combinations of conditions. Clearly the out-of-sample performance of the networks has strongly indicated their relevance in personnel research and modeling. Perhaps more important is the insight that can be gained into decision making and other processes as demonstrated by the response surfaces for the aggregate reenlistment rate. In lieu of the constant impact or constant elasticity of most regression methods, a successfully trained network offers more insight into the structure of the problem. For example, the effect of a change in the unemployment rate depends on the current rate; or, assumptions about the impact of a change in military compensation must be made in the context of the current unemployment rate. While the wealth of information available from a rich model such as one developed by a neural network can be difficult to analyze, ignoring and obscuring important features by forcing them to fit a preconceived functional relation seems more dangerous. With the proper tools, the interrelation and features developed by a network can be made available as a more realistic model of the process being analyzed. This task merely served as a test-bed for approaching such problems as rate projection, decision making, and selection in the Air Force personnel context. In at least 2 of the areas examined, networks have proven to be ready for

more extensive application. Many personnel management tasks and problems can be approached using the tools tested in this research.

Several of the methods employed in the current research were developed and implemented in software specifically for this task. In particular, the refinements to prevent over-training of neural networks are not currently available in commercial neural network software. For neural networks to be useful in the personnel context, easily used systems to develop networks using the procedures outlined in this document must be implemented. Additional tools are also required to elucidate the relationships developed by a trained neural network model. Contrary to popular opinion, the impact of input factors in a neural network model can be analyzed; as shown in Figures 8 through 11. However, for this process to be widely applicable, methods of automatically exploring the response surface of a network must be made available. With the development of these 2 tools, neural networks should become widely applicable in personnel research.

REFERENCES

Ash, T. (1989). *Dynamic node creation in back propagatic networks* (ICS Report 8901). La Jolla, CA: Institute for Cognitive Science.

Ash, C., Udis, B., & McNown, R.F. (1983). Enlistments in the all-volunteer force: A military personnel supply model and its forecasts. *American Economic Review, 73*(1), 145-155.

Atlas, L., Cole, R., Conner, T., El-Sharkawi, M., Marks, R.J., II, Muthusamy, Y, & Barnard, E. (1990). Performance comparisons between back propagation networks and classification trees on three real-world applications. *Neural Information Processing Systems 2,* Touretski, D.S. (ed.), San Mateo, CA: Morgan Kaufmann Publishers, 623-629.

Cacoullos, T. (1966). "Estimation of a Multivariate Density," *Annals of the Institute of Statistical Mathematics (Tokyo),* vols. 18, num. 2, pp 179-89, 1966.

Carpenter, M.A., Monaco, S.J., O'Mara, F.E., & Teachout, M.S. (1989). *Time to job proficiency: a preliminary investigation of the effects of aptitude and experience on productive capacity* (AFHRL-TP-88-17, AD-210 575). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

Carter, G.M., Murray, M.P., Arguden, R.Y., Brauner, M.K., Abrahamse, A.F., Greenberg, H., & Skoller, D.L. (1987). *Middle-term loss prediction models for the Air Forces' enlisted force management system.* (R3482-AF). Santa Monica, CA: The RAND Corporation.

Carter, G.M., Skoller, D.L., Perrin, S.E., & Sakai, C.S. (1988). *An enlisted force management system to predict the effects of bonus decision.* Santa Monica, CA: The RAND Corporation.

Computing Resource Center (1989). *Stata Reference Manual, Release 2*, Computing Resource Center, Los Angeles, CA.

Cowan, J.D., & Sharp, D.H. (1988). Neural nets and artificial intelligence. *The Artificial Intelligence Debate*, Graubard, Stephen R. (ed.), Cambridge, MA; London, England: The MIT Press, 85-121.

Denker, J.S., Gardner, W.R., Graf, H.P., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D., Baird, H.S., & Guyon, I. (1989). Neural network recognizer for hand-written zip code digits. *Neural Information Processing Systems 2,* Touretski, D.S. (ed.), San Mateo, CA: Morgan Kaufmann Publishers, 323-331.

DeVany, A.S., & Saving, T.R. (1982). Life-cycle job choice and the demand and supply of entry level jobs: Some evidence from the Air Force. *The Review of Economics and Statistics*, LXIV(3), 457-465.

DeVany, A.S., Saving, T.R., & Shughart, W.F. (1978). *Supply and equlibrium inventory of Air Force enlisted personnel: A simultaneous model of the accession and retention markets incorporating force level constraints.* (AFHRL-TR-78-10, AD-A058 097) Brooks AFB, TX: Occupation and Manpower Research Division,, Air Force Human Resources Laboratory.

Duda, R., & Hart, P. (1973). *Pattern Classification and Scene Analysis*, John Wiley and Sons, NY.

Fahlman, S.H. (1988). Faster-learning variations on back-propagation learning with second order methods. *Proceedings of the 1988 Connectionist Models Summer School, June 17-26, 1988, Carnegie Mellon University*, Touretski, D.S., Hinton, G., & Sejnowski, T. (eds.), San Mateo, CA: Morgan Kaufmann Publishers, 133.

Faneuff, R.S., Valentine, L.D., Stone, B.M., Curry, G.L., & Hageman, D.C. (1990). *Extending the time to proficiency model for simultaneous application to multiple jobs* (AFHRL-TP-90-42, AD-A224 759). Brooks AFB: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Funahashi, K. (1989). On the approximate realization of continuous mappings by Neural Network. *Neural Networks, 2*(3), 183-192.

Gorman, R.P., & Sejnowski, T.J. (1988). Analysis of hidden units in a layered network trained to classify sonar targets. Neural Networks, *1*(1), 75-89.

Hecht-Nielson, R. (1987). Kolmogorov mapping neural network existence theorem. *IEEE First International Conference on Neural Networks, San Diego, CA, June 21-24, 1987, 3,* 11-13.

Hedge, J.W., & Teachout, M.S. (1986). *Job performance measurement: A systematic program of research and development* (AFHRL-TP-86-37, AD-A174 175). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

Hornik, K., Stinchcomebe, M., & White, H. (1989). Multilayer feed forward networks are universal approximators. *Neural Networks, 2*(5), 359-366.

Josin, G. (1990). Development of a neural network autopilot model ford high performance aircraft. *International Joint Conference on Neural Networks, January 15-19, 1990, Washington D.C., IEEE & INNS, 1,* Caudill, M (ed.), II - 547.

Karnin, E.D. (1990). A simple procedure for pruning back-propagation trained neural networks. *Transactions on Neural Networks*, *1*(2), 239-242.

Kimoto, T., Asakawa, K., Yoda, M., & Takeoka, M. (1990). Stock market prediction with modular neural networks. *International Joint Conference on Neural Networks, San Diego, CA, June 17-21, 1990, IEEE & INNS, 1*, I - 1-6.

Klimasauskas, C.C. "Casey" (1988). Neural networks: A short course from theory to application. *PC AI*, November/December 1988, 26-30.

Kmenta, J. (1971). *Elements of Econometrics*, Macmillan Publishing Co., NY.

Kohler, D.F. (1988). *Using survivor functions to estimate occupation-specific bonus effect.* (R-3348-FMP). Santa Monica, CA: The RAND Corporation.

Kohonen, T. (1984). *Self-organization and associative memory* (3rd ed.). New York: Springer-Verlag.

Koutsoyiannis, A. (1977). *Theory of Econometrics, Second edition*, Macmillan Press, London.

Ladd, G. (1986). "Linear Probability Functions and Discriminant Functions," *Econometrica*, vol. 34, pp 873-85.

Lakhani, H., Gilroy, C., & Capps, C. (1984). *Logistic microdata model of first term Army reenlistment.* (MPPRG working paper 84-2). Alexandria, VA: Manpower and Personnel Research Laboratory, U.S. Army Research Institute for the Behavioral and Social Sciences.

Lakhani, H. (1987). *Bonuses, training costs, and quit rates: A three stage least squares approach.* (TR 734). Alexandria, VA: Manpower and Personnel Research Laboratory, U.S. Army Research Institute for the Behavioral and Social Sciences.

Leamer, E.D. (1978). *Specification Searches.* John Wiley & Sons, New York - Chichester - Brisbane - Toronto.

Leung, H.C., & Zue, V.W. (1988). Applications of error back propagation to phonetic classification. *Neural Information Processing Systems 1*, Touretski, D.S. (ed.), San Mateo, CA: Morgan Kaufmann Publishers, 206-214.

Maddala, G. (1985). *Limited-dependent and qualitative variables in econometrics*, Cambridge University Press, NY.

Michelson, W.T., & Rydell, P.C. (1989). *Aggregate dynamic model (ADAM) for Air Force enlisted personnel: Technical documentation.* (N-3020/2-AF). Santa Monica, CA: The RAND Corporation.

Moody, J., & Darken, C. (1988). Learning with localized receptive fields. *Proceedings of the 1988 Connectionist Models Summer School, June 17-26, 1988, Carnegie Mellon University*, Touretski, D.S., Hinton, G., & Sejnowski, T. (eds.), San Mateo, CA: Morgan Kaufmann Publishers, 133.

Morgan, N., & Bourlard, H. (1990). Generalization and parameter estimation in feed forward nets: some experiments. *Neural Information Processing Systems 2*, Touretski, D.S. (ed.), San Mateo, CA: Morgan Kaufmann Publishers, 630-637.

Mozer, M.C., & Smolensky, P. (1989). "Skeletonization: A technique for trimming the fat from a network via relevance assessment," in *Advances in Neural Information Processing I*, D. S. Touretski, Ed. Morgan Kaufmann, pp. 107-115.

Odom, M.D., & Sharda, R. (1990). A neural network model for bankruptcy prediction. *International Joint Conference on Neural Networks, San Diego, CA, June 17-21, 1990, IEEE & INNS, 2*, II - 163-168.

Parzen, E. (1962). "On Estimation of a Probability Density Function and Mode," *Annals of Mathematical Statistics*, vol. 33, pp. 1065-76.

Rumelhart, D.E. (1990). Brain style computation: Neural networks and connectionist AI (oral presentation). Las Vegas TIMS/ORSA Joint National Meeting, May 7-9, 1990.

Rumelhart, D.E., & McClelland, J.L. (eds.) (1986). *Parallel Distributed Processing, Explorations in the Microstructure of Cognitions, Volume 1 & 2.* Cambridge, MA: The MIT Press.

Saving, T.R., & Stone, B.M. (1982). *Analysis of the impact of "people programs" upon retention of enlisted personnel in the Air Force.* (F41689-81-C-0063). Randolph AFB, TX: AFMPC/MPCYA, Final Report.

Saving, T.R., Stone, B.M., Looper, L.T., & Taylor, J.T. (1985). *Retention of Air Force enlisted personnel: An empirical examination.* (AFHRL-TP-85-6). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Sietsma, J. & Dow, R.J.F. (1988). Neural net pruning - why and how. *IEEE International Conference on Neural Networks, San Diego, CA, July, 1988, 1*, I - 325-333.

Shea, P. M., & V. Lin (1989). "Detection of Explosives in Checked Airline Baggage Using and Artificial Neural System," *Proceedings of the International Joint Conference on Neural Networks, Washington D.C., June 18-22, 1989*, vol II, pp. 31-41, IEEE TAB Neural Network Committee, SOS Printing, San Diego, CA.

Smith, A.D., Sylvester, S., & Villa, C. (1989). *Army reenlistment models.* (MDA903-87-C-0590). U.S. Army Research Institute.

Specht, D.F. (1990). Probabilistic neural networks. *Neural Networks, 3*(1), 109-118.

Specht, D.F. (1988). Probabilistic neural networks for classification, mapping, or associative memory. *IEEE International Conference on Neural Networks, San Diego, CA, July, 1988, 1*, I - 525-532.

Steel, G., & Torrie, J. (1960). *Principles and Procedures of Statistics*, McGraw-Hill, New York.

Stone, B.M., Looper, L.T., & McGarrity, J.P. (1990). *Validation and reestimation of an Air Force reenlistment analysis model.* (AFHRL-TP-89-55, AD-A218 794). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Stone, B.M., Saving, T.R., Turner, K.L., Looper, L.T., & Engquist, S.K. (1991). *A simultaneous estimation model of Air Force accession and retention* (AL-TR-1991-0063, AD-A240 064). Brooks AFB, TX: Human Resources Directorate, Manpower and Personnel Division, Armstrong Laboratory.

Stone, B.M., Turner, K.L., Fast, J.C., Curry, G.L., Looper, L.T., & Engquist, S.K. (1991). *A computer simulation modeling approach to estimating utility in several Air Force specialties* (AL Technical Paper in review). Brooks AFB, TX: Human Resources Directorate, Manpower and Personnel Division, Armstrong Laboratory.

Stone, B.M., Wortman, R.A., & Looper, L.T. (1989). *Air force retention analysis package: users manual.* (AFHRL-TP-89-10, AD-A212 767). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Surkan, A.J., & Singleton, J.C. (1990). Neural networks for bond rating improved by multiple hidden layers. *International Joint Conference on Neural Networks, San Diego, CA, June 17-21, 1990, IEEE & INNS, 2*, II - 157-162.

Syllogistics Inc., & RRC, Inc. (1989). *Epps VI.0 design manual.* Springfield VA: Syllogistics Inc., for the Analysis Division and Planning Division, Directorate of Personnel Plans, Washington, D.C.

Terza, J.V., & Warren, R.S. (1986). *U.S. Army reenlistment and extension by occupation: A reduced-form trinomial probit approach. (DAAG29-81-D-0100, Delivery Order 1748). Alexandra, VA: U.S. Army Research Institute.*

Waibel, A.H. (1989). Consonant recognition by modular construction of large phonetic time-delay neural networks. *Neural Information Processing Systems 1*, Touretski, D.S. (ed.), San Mateo, CA: Morgan Kaufmann Publishers, 215-223.

Warner, J.T., & Goldberg, M.S. (1983). The influence of non-pecuniary factors on labor supply: The case of Navy enlisted personnel. *The Review of Economics and Statistics,* XX(Y), 26-35.

Werbos, P. (1974). *Beyond regression: New tools for prediction and analysis in the behavioral sciences* (Ph.D. thesis). Cambridge, MA: Harvard University Committee on Applied Mathematics.

Wiggins, V.L. (1990). *Neural network applications literature review.* Informal technical report for Air Forces Human Resources Laboratory, Brooks AFB, TX, Contract No. F41689-88-D-0251, Task 29. Bryan, TX: RRC, Inc.

Wiggins, V.L., Looper, L. T., & Engquist, S.K. (1991a). *Neural networks: A primer* (AL-TP-1991-0011, AD-A235 920). Brooks AFB, TX: Human Resources Directorate, Manpower and Personnel Division, Armstrong Laboratory.

Wiggins, V., Looper, L. T., & Engquist, S.K. (1991b). *Neural networks and their application to Air Force personnel modeling* (AL-TR-1991-0031, AD-A243 802). Brooks AFB, TX: Human Resources Directorate, Manpower and Personnel Division, Armstrong Laboratory.